IDENTIFYING FUNCTIONAL ADAPTATIONS ASSOCIATED WITH PATHOGENICITY IN BACTERIA

A thesis submitted in partial fulfilment of the requirements for the Degree of Doctor of Philosophy in Biochemistry

in the University of Canterbury

by Nicole E. Wheeler

University of Canterbury

2017

Table of contents

Acknowledgements

I would like to thank my supervisor, Paul Gardner, for providing me with the opportunities and mentoring I needed to embark on this project. I appreciate the freedom I was afforded to pursue my own interests and ideas, and the patience I was given when I was stuck in unfamiliar territory. The work presented in this thesis would not be what it is today without the influence of the people I met and the ideas I was exposed to as a result of the countless opportunities I was given to travel and present my own research. Finally, I would like to thank him for teaching me not to take myself so seriously.

I thank my employer, Anna Pilbrow, for giving me an opportunity to learn on the job and providing me with endless encouragement. Her wisdom and rigour when it came to designing and communicating research helped me to improve my own abilities as a scientist, and the ideas and skills I learned on the job have proven invaluable to the development of this body of work.

I thank my collaborators for putting their trust in me and giving me the opportunity to try something different and relatively untested with their data. I appreciate the feedback I was provided and the domain knowledge I picked up along the way. I would like to thank all of the incredible scientists that I have met during this process, including those who made time to meet with me despite having never heard of me before, for sharing their ideas and being open to mine.

I thank my friends for providing an escape from work and comic relief when I needed it. I also thank them for being understanding when I wasn't available as much as they wanted. They have been supportive through my successes and struggles, and helped me to maintain perspective.

I would like to say thank you to my family, who have always supported my education and my interests. Thank you to my mother, for inspiring my interest in health and genetics, and my father for inspiring my interest in informatics. Thank you to both of my parents for teaching me the value of critical thought, and of being true to my values and my opinions. And finally, thank you for being proud of the path I chose to take.

Abstract

Next generation sequencing technologies have provided us with a wealth of information on genetic variation, but predicting the functional significance of this variation is a difficult task. This thesis summarises the development of a profile hidden Markov model based approach we call delta-bitscore (DBS). The approach identifies orthologous proteins that have diverged at the amino acid sequence level in a way that is likely to impact biological function. I present the benchmarking of this approach using several widely used datasets and its application to various biological questions. I then outline the extension of this method from a pairwise comparative method to one that can be scaled for the comparison of hundreds or thousands of bacterial genomes. I demonstrate the utility of the method for identifying associations between genetic variation and phenotypes of interest, and discuss methodological considerations and extensions that must be made in order for this approach to function effectively on a large scale.

List of abbreviations used

AUC	Area under the curve
c-di-GMP	Cyclic di-GMP
COG	Clusters of orthologous groups
CoNS	Coagulase negative staphylococci
DBS	Delta-bitscore
GWAS	Genome-wide association study
HGT	Horizontal gene transfer
LD	Linkage disequilibrium
LOF	Loss of function
OD	Optical density
OOB	Out-of-bag
OXC	Oxfordshire Surveillance Project
PCA	Principal component analysis
ROC	Receiver operating characteristic
SNP	Single nucleotide polymorphism
T3SS	Type 3 secretion system
VI	Variable importance

Chapter One | Introduction

Context

Historically, contagious pathogenic bacteria have devastated communities globally (Fauci & Morens, 2012), and they remain a significant cause of morbidity and mortality worldwide (World Health Organization, 2013). Infectious disease differs from other human illnesses in its potential for unpredictable and explosive spread (Fauci & Morens, 2012). With globalization and population density increasing over time, infectious disease can spread more quickly, making the recognition and prevention of outbreaks of infectious bacteria a serious consideration for health authorities (Morens & Fauci, 2013). Epidemiological data in New Zealand stresses the need for improved infection prevention measures. In a national study of New Zealand hospital admissions from 1989 to 2008, infectious diseases made the largest contribution to acute hospital admissions, and they showed a striking increase over time compared to non-infectious diseases. Camplylobacteriosis, acute rheumatic fever, childhood pneumonia, meningococcal disease and skin infections are all unusually prevalent in New Zealand, indicating that our efforts to combat infectious disease are falling behind those of other countries (Baker et al., 2012).

Antibiotic resistant pathogens in particular are an increasing concern, with the WHO estimating that the rate of fatal multi-drug resistant bacterial infections could increase from 700,000 in 2014 to 10 million by 2050 (O'Neill, 2016), and one of the first cases of a death due to a bacterial pathogen that was resistant to all available antibiotics occurring early this year (Chen et al., 2017). This case also highlights the increased threat of pandemic outbreaks due to globalization, as the patient acquired the infection in India, then brought it home to the United States. Many of these infections can be asymptomatic, and as a result, they could become globally widespread prior to detection in the clinic.

The prevention of outbreaks of pathogenic bacteria relies on early warning, as failure to detect pathogenic microbes before their populations have reached high numbers and become widely dispersed makes effective elimination by antimicrobial intervention more difficult (Christaki, 2015). As the cost of genome sequencing decreases (van Nimwegen et al., 2016), and sequencing technologies become more portable (Quick et al., 2016), the option of monitoring pathogens in the environment using next-generation sequencing becomes more realistic. Genomics-based pathogen sensors offer advantages over more traditional methods in terms of speed, sensitivity, and specificity (Law et al., 2014). However,

these approaches can only identify known pathogens or known combinations of virulence factors, which limits their ability to predict outbreaks of disease caused by novel pathogens. This is not a limitation of genomics-based pathogen sensing technologies alone, but rather one imposed by most currently used techniques that do not employ predictive methods to detect novel pathogens (Christaki, 2015).

Due to the diversity of bacterial pathogens and their ability to constantly adapt, it is important to have a sensor that can detect multiple pathogens, and can identify genetic markers that provide information on the nature of the pathogen, such as antibiotic resistance and toxin genes. A detector should also be able to distinguish between pathogens and closely related nonpathogenic species. This can be a difficult task, as closely related nonpathogens can harbour similar repertoires of virulence factors and other markers of pathogenicity (Barbosa et al., 2014). Identifying signatures of an early transition from a nonpathogenic to a pathogenic lifestyle would constitute a major advance in our ability to detect novel pathogens.

Bacterial pathogen evolution through loss of function

Before we had access to the wealth of genomic data available today, the scientific community initially hypothesised that the reason some bacteria were pathogenic was because they had acquired virulence factors, such as genes encoding toxins (Wu et al., 2008). However, as the genomes of more bacterial pathogens and closely related nonpathogenic bacteria became available, a different picture was revealed, prompting a reconsideration of what constitutes a bacterial pathogen. In contrast to the expectation that pathogen genomes would be larger due to greater acquisition of genes to assist with infection, their genomes were typically smaller (Merhej et al., 2013). Closely related nonpathogenic bacteria were found to also carry virulence factors, as well as antivirulence factors that were not present in pathogens. In addition, the genomes of pathogenic bacteria were characterised by a larger number of pseudogenes, fewer transcriptional regulators, disordered operons, and more genes encoding DNA replication and repair proteins (Merhej et al., 2013), suggesting a mutational burden on these pathogens due to reduced effectiveness of selection. In fact, large-scale comparative genomic analysis of pathogenic and nonpathogenic bacteria has shown that gene loss appears to be a more significant factor in the evolution of pathogenesis than the presence of virulence factors (Georgiades & Raoult, 2011b).

Bacterial genomes have high mutation rates, with a strong bias toward deletions (Mira et al., 2001). As a result, genes not maintained by selection are expected to undergo inactivation over time. The rate of loss of genes will be determined partly by selection pressure to maintain their function, and partly by the efficiency of selection. A key distinction to make in characterising gene loss associated with pathogenicity is whether the loss was adaptive or due to drift (Koonin, 2016). Gene loss in bacterial pathogens will occur as a result of both of these processes, and distinguishing the underlying cause of a loss of function can be key to determining its relevance to pathogenicity.

Gene loss may be the result of adaptive mutations that facilitate a pathogenic lifestyle (pathoadaptation). If a bacterium encounters a new host environment and is incidentally able to persist and grow, its fitness in the new environment is likely to be suboptimal. Genes required for fitness in the old environment may actually decrease fitness in the new environment, where a new set of selective pressures will likely be present. Adaptation to a new pathogenic niche can sometimes be achieved by selective inactivation of ancestral genes that are no longer compatible with the environment that the bacteria are living in. In fact, under severely nutrient limiting conditions, a substantial improvement in fitness can be achieved purely by loss of function mutations (Hottes et al., 2013). It can be difficult to distinguish pathoadaptive mutations from genome reduction computationally, however, if a trait is absent in independently evolved pathogenic clones of a species, but expressed in closely related nonpathogenic species, this is thought to be strongly suggestive of a pathoadaptive mutation that has arisen by convergent evolution. Experimental validation to confirm that expression of the gene in a pathogenic clone attenuates virulence is important for achieving greater confidence in these predictions (Maurelli, 2007).

Many gene loss events observed during pathogen evolution appear to be very similar to those seen in host adaptation. This can be explained by the fact that both adaptations involve specialization to a similar niche (Merhej et al., 2009), and that many pathogenic bacteria evolve from host-adapted species, meaning that many of the gene losses have occurred prior to the development of a pathogenic phenotype (Pallen & Wren, 2007). While many of these patterns of gene loss represent the shedding of functions that are no longer required due to the nutrient richness and greater stability of the host environment (Medina & Sachs, 2010), a proportion of gene losses are also expected to be due to a reduced effective population size in host adapted bacteria resulting in an increase in genetic drift. Genetic isolation, rather than parasitism, is thought to have a greater impact on genome reduction,

as it reduces the ability of negative selection to maintain gene function across the genome, rather than relaxing selective constraints on specific genes. This effect becomes particularly pronounced if isolation leads to degradation of recombination and repair machinery, leading to accelerated accumulation of mutations (Georgiades & Raoult, 2011a).

Comparative genomics has revealed bacterial pseudogenes to be more common than we first thought. We have discovered that they can be incredibly useful to us in our quest to better understand the molecular changes that bacteria undergo as they adapt to new niches. Pseudogenes give us a glimpse back in time, by showing us an indication of functions that bacteria used to perform, that they no longer can (Goodhead & Darby, 2015). Thus, a major focus of this thesis is on identifying genes that have recently accumulated mutations that disrupt function, and identifying associations between these loss of function events and adaptation to a pathogenic niche.

Predicting the effects of sequence variation

A number of previous studies have inferred the loss of genes in pathogenic bacteria based on their presence in nonpathogenic close relatives (Georgiades & Raoult, 2011b; Merhej et al., 2013; Pallen & Wren, 2007). However, losses of gene function due to inactivating mutations would provide a more sensitive signal of the change in the genomes of pathogens than complete loss of the gene by deletion, as there would be a larger number of loss of function events to measure, and we would form a more comprehensive picture of functional losses in pathogens than if we looked at deletions alone. A measure of the functional disruption of genes could provide a better early indicator of characteristic shifts in the coding potential of a genome that may be associated with the development of a pathogenic phenotype, due to the fact that it would pick up more subtle changes in the genome that often precede complete gene loss (Kuo & Ochman, 2010).

Studies have previously incorporated information on pseudogenization into the study of bacterial pathogen evolution, with results that yielded great insights into the functional requirements of a pathogenic lifestyle that could not have been gained through looking at gene loss and gain alone (Nuccio & Baumler, 2014; Parkhill et al., 2003; Rohmer et al., 2007). However, these approaches typically rely on manual annotation of pseudogenes based on large truncations, frameshifts or indels within the coding sequence. This process is time-consuming and usually limits investigations to a small number of genomes. In addition, this limits the identification of functional losses to those most obvious changes that can be confidently identified with the naked eye.

A predictive model for determining whether a sequence mutation is deleterious to gene function would improve the rate and ease of identification of bacterial pseudogenes, and would allow this approach to become more commonplace in large scale comparative genomics studies. This type of model would be simplest to develop for protein coding genes, as it is easier to anticipate the consequences of mutations in protein sequences, given information on the evolutionary conservation across the sequence and the biochemical similarities and differences between amino acids.

A number of approaches that can be applied to bacterial genomes already exist for the detection of nonsynonymous mutations that would disrupt protein function (see Chapter Two for an outline of comparable approaches). Most of these methods have been primarily geared toward the analysis of human sequences (Adzhubei et al., 2013; Choi & Chan, 2015; Reva et al., 2011), with many providing pre-computed values for many human proteins available for download, bypassing the need to run the same analyses repeatedly. We did not anticipate that these methods would scale well for large comparative bacterial genomics projects, due to the fact that many of them need individual mutations to be coded according to a convention accepted by the method, which often only allows assessment of a single change per protein (Adzhubei et al., 2013; Choi et al., 2012; Reva et al., 2011). Thus, during the course of this thesis I have developed and tested a profile hidden Markov model-based approach to measure the impact of sequence variation on protein function. The method was originally developed for pairwise comparisons (see Chapters Two and Three), and subsequently adapted to facilitate comparisons of multiple genomes (see Chapters Four, Five and Six).

Most methods for predicting the impact of sequence variation use a BLAST-based approach to assess which mutations will result in loss of protein function and which will be well tolerated (Adzhubei et al., 2013; Choi et al., 2012; Ng & Henikoff, 2006; Reva et al., 2011). Profile hidden Markov models present a potentially superior approach to typical BLAST-based approaches for predicting the effects of sequence variation because the model explicitly penalises mutations in highly conserved sites more than highly variable sites (Shihab et al., 2013). Put briefly, profile hidden Markov models are statistical models built from alignments of homologous sequences (or sometimes single sequences). For each column in the alignment, the model captures the frequencies of each amino acid that occurs, as well as the frequencies of insertions and deletions (Eddy, 1998). The utilisation of prior

information, containing background frequencies of residues in the form of Dirichlet mixture, is an approach which prevents the over-training of the model, particularly in instances where the alignment has low sequence diversity, and improves the overall generalizability of the model (Sjolander et al., 1995). Dirichlet mixture densities are calculated by analyzing the frequencies of amino acids at each position in a large collection of sequence alignments. High probabilities are given to commonly occurring combinations of amino acids, and low probabilities to combinations of amino acids that are not frequently seen together in the same column of an alignment (Sjolander et al., 1995).

Several methods have already been proposed that use profile hidden Markov models to predict the consequences of mutations in protein coding genes. The first was logR.E, a method that was proposed in 2004 (Clifford et al., 2004), which was subsequently found to perform poorly compared to other methods in predicting cancer causing mutations (Gnad et al., 2013), likely due to the use of profile hidden Markov models (HMMs) that were too general for the task of detecting deleterious mutations (Clifford et al., 2004). Another method, called FATHMM, was primarily designed to detect deleterious mutations in human genes, but offers an implementation of their approach that is species-independent, however this version also has little power to discriminate between pathogenic variants and nonpathogenic variants and performs worse than BLAST-based methods (Shihab et al., 2013).

Because the performance of both of these methods was so poor, we developed our own approach, presented in Chapter Two. The approach is based on a similar premise, but uses profile HMMs with higher sequence identity to the query protein than those methods typically used for distant homology searching, such as Pfam and SUPERFAMILY models (Gough & Chothia, 2002; Sonnhammer et al., 1997). One of the primary objectives of our newly developed method was to identify functionally significant changes in protein coding genes that were associated with niche adaptation and pathogenicity in bacteria. To do this, we needed to design a statistical framework with the ability to identify these associations. This thesis presents a combination of traditional nonparametric statistical approaches and machine learning approaches to identify these associations, with special consideration to the scalability of the analysis in anticipation of the need to analyze hundreds or thousands of bacterial genomes (Field et al., 2006).

Gaining insight from massive datasets

Classical statistical methods provide a range of techniques to assess the relationship between a potentially predictive variable and an outcome. These techniques have provided us with medically important insights in the past, however in an age of big data and cheap sequencing, we are seeing more situations where classical statistical techniques are inappropriate due to the sheer volume of data (Babyak, 2004). A key issue that has arisen is known as the 'big p, little n' problem ($p >> n$), in which an investigation has vastly more variables than samples, leading to a large multiple testing burden which is difficult to overcome using traditional statistical techniques. These problems abound in bioinformatics, where although tens or thousands of samples in an experiment are common, these yield thousands to millions of potential variables, such as genetic variants, gene expression values, etc. Using many univariate tests for significant associations results in many true positive associations being lost during correction for multiple testing, whereas using multivariate regression modelling of an outcome will use up too many degrees of freedom and result in overfitting of the model to the data (Babyak, 2004). This issue has prompted much discussion of the best way to deal with multiple testing and overfitting issues, both in bioinformatics and in the broader field of big data (Liao & Chin, 2007; Loh, 2012; Zhang et al., 2008).

However, in spite of the statistical challenges before us, a more optimistic view of the dramatic increase in data available to us has also been put forward. Some argue that this new deluge of data marks the end of the need for scientists to formulate hypotheses and test them, that we should instead let the data speak for itself. More specifically, they argue that mining the patterns and relationships captured in these large datasets using sophisticated algorithms can organically produce meaning and insight about complex processes that could not have been captured by a traditional scientific approach, in essence 'revealing patterns that we didn't even know to look for' (Kitchin, 2014).

While this sounds promising, our understanding of how to capture meaningful information from big data is still in its infancy, as demonstrated by the Google Flu Trends (GFT) predictor (Lazer et al., 2014), which consistently over-predicted flu trends, generating results similar to those that could be obtained by simply extrapolating Centers for Disease Control (CDC) reporting trends from the weeks prior. A major criticism of GFT was that the model used to predict future flu cases was opaque, giving little insight into the workings of the model that was making these predictions, and which variables it was relying on. Intuitively, we would

expect the model to be heavily reliant on flu-related search terms, but in fact the model appeared to be heavily overfit to seasonal terms unrelated to illness. This revealed a critical vulnerability in the model - with so many variables, it was easy to overfit the model, allowing perfect prediction accuracy on training data but failing to capture the true causative factors behind the phenomenon they were interested in. A key disconnect between the values of the developers of many of these approaches, and our interest in them as biologists can be illustrated by a quote from Eric Siegel in his book Predictive Analytics:

> *'We usually don't know about causation, and we often don't necessarily care...the objective is more to predict than it is to understand the world...It just needs to work; prediction trumps explanation'.*
> - *(Siegel, 2013), quoted in (Kitchin, 2014)*

This argument for hypothesis free, opaque predictive software is made largely by the business sector (Siegel, 2013), where if a recommendation by a model increases profits, the underlying phenomenon causing the effect is considered inconsequential. However in biology, often the predictive power of the model is of secondary concern, compared to the insights that can be gained by uncovering the underlying processes generating the data. This discrepancy in philosophy and objectives creates a need for the adoption or development of modelling approaches geared more toward data analysis for understanding rather than performance.

With such a large amount of biological sequence data becoming available, and potential complexity in the way variables interact, adoption of more sophisticated algorithms for identifying patterns in biological data offers us a chance to better capture complex evolutionary phenomena, like adaptation to a niche or an increase in virulence. Machine learning techniques offer an opportunity to start to capture some of these higher level interactions (Bureau et al., 2005; Motsinger & Ritchie, 2006; Yoon et al., 2003). Machine learning generally refers to the use of algorithms to identify patterns in data for prediction and classification purposes, trained using a dataset thought to be representative of a broader real life situation (Tarca et al., 2007).

A variety of machine learning algorithms have been developed, each with their strengths and weaknesses. Of the various approaches, random forests are gaining popularity in the life sciences for a number of reasons (Touw et al., 2013). Primarily, random forests behave less

like a ‘black box’ than other methods - an interested party can dissect each decision made by the model, and the contributions made by each variable to the overall decision process. Random forests are also explicitly designed to prevent overtraining of a model, even in cases where p >> n. Finally, random forests tend to perform better than other methods in classification and regression problems (Caruana et al., 2008).

The random forest approach was an ensemble learning method originally developed by Leo Breiman (Breiman, 2001). Ensemble methods combine a set of weakly predictive models, to form a high performing composite model. In the case of random forests, the ensemble consists of decision trees each built using subsampling from the larger dataset. This subsampling during training improves the generalizability of the model and prevents overtraining of the model on the data (Breiman, 2001). It also allows the training process to proceed much faster than if all variables were considered for all trees, enhancing its ability to scale to large datasets.

Growth of a single tree involves randomly sampling some of the training samples with replacement. Within the tree, each node is chosen by randomly sampling some of the variables with replacement and selecting the one that best splits the data into two more homogenous groups (Breiman, 2001). The tree is grown until each terminal node has a certain, pre-specified number of observations in it (typically 5 for regression and 1 for classification (Liaw & Wiener, 2002)). A prediction is made by the model by combining the predictions made by each individual tree, using a voting system in the case of classification problems (Breiman, 2001).

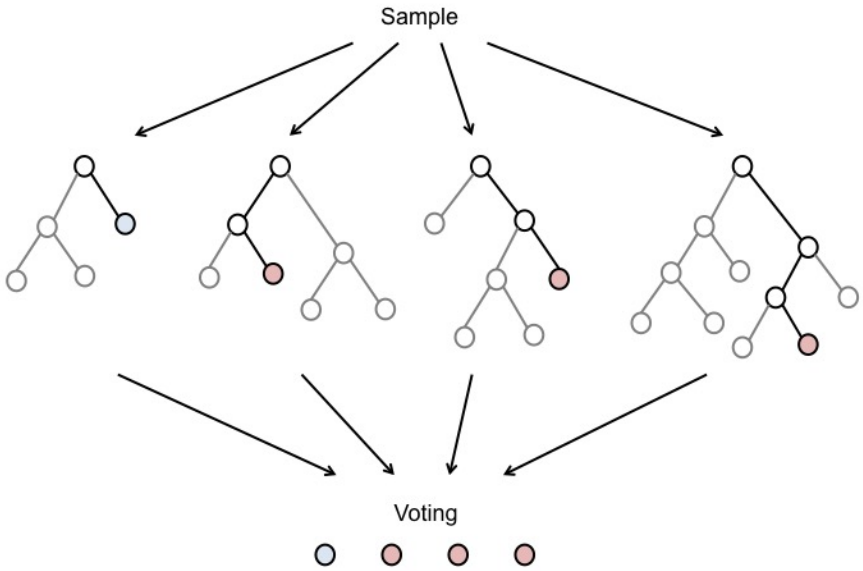


Figure 1 | A schematic representing the functioning of a random forest classifier
Each sample is run through each tree in the forest. It flows down the tree based on whether it meets a criterion presented by each node. Upon reaching a terminal node, a classification will be made, then classifications made by each tree are collated and a decision is made using a voting process.

Prediction error of the model is calculated during the process of building the model. For each individual tree, the samples taken to build trees are used as training data, and the samples that have been left out (out-of-bag (OOB) samples) are then run through the tree as testing data. The accuracy of the tree in predicting OOB samples is then recorded as OOB error (Breiman, 1996). Each tree is essentially naive to its OOB test data, so prediction accuracy can be interpreted as ability of the model to make general predictions on previously unseen data, provided the training data is a fair representation of real world data.

Feature selection can be used to reduce the number of variables that go into training a model. Reasons for doing this include increasing the speed of model training, making the resulting model more interpretable and improving prediction accuracy (Haury et al., 2011). The simplest form of feature selection is to remove variables that do not vary across the training set, as these will offer no possibility of an informative split, but will still be sampled in place of informative variables. In many datasets this is an unnecessary step but when dealing with genomic data in closely related isolates there is the possibility of having genes whose sequences are identical across all samples. Pre-filtering methods such as a t-test can also be used to get a rough indication of how well each feature relates to the class being predicted. Variables with high p-values can then be excluded from model building, or

variables can be weighted based on how much they are expected to contribute to prediction accuracy (Amaratunga et al., 2008). Alternatively, a random forest can be built using all variables, then the variables with higher feature importance can be used to build a new, more accurate model (Goldstein et al., 2010).

Once a random forest is built to predict a phenotype of interest using genomic data, it can then be further explored to gain additional insight into the data. For each variable in the model, a measure of variable importance is computed. This can either be measured as Gini index (the increase in purity of daughter nodes achieved by splitting data based on a given variable) or as the decrease in prediction accuracy caused by permutation of the values of a variable. Variable importance measures are computed for each tree, then the average is taken to give an indication of the overall contribution of each variable to prediction accuracy. These variable importance measures can then be used to rank genes based on how likely they are to be relevant to a phenotype of interest. The use of these variable importance measures has been a focus of this thesis, however random forests can be further probed for information on the dataset that may be relevant to an evolutionary process of interest (see Discussion, Touw et al., 2013).

This thesis explores the utility of the random forest method for identifying genes that are informative of phenotype. Random forests are already being explored for their value in identifying associations between single nucleotide polymorphisms (SNPs) and diseases in humans (Goldstein et al., 2010), and in bacteria (Dutilh et al., 2014), however the pairing of the delta-bitscore metric developed in this thesis and random forests offers improvements over a SNP-based approach in terms of reducing the sparsity of datasets (i.e. the number of variables that are offer little or no predictive value) and reducing the number of variables incorporated into the model.

Thesis outline

The following chapters outline the development of the delta-bitscore method of predicting functional disruption of protein coding sequences, from pairwise comparisons of bacterial proteomes, to a highly scaleable approach designed with large comparative genomics studies in mind. The thesis finishes with a chapter considering phenotype arrays, another source of a vast amount of data that could be used to characterise niche adaptation, with considerations of how to scale analysis of this form of data and eventually integrate it with delta-bitscore analyses. The published and draft manuscripts in this thesis proceed as follows:

Chapter Two is a published paper outlining the development, testing and use of our original profile hidden Markov model based approach to measuring the impact of nonsynonymous mutations on protein function. This application focuses on the pairwise comparison of proteomes to identify those proteins shared by both organisms that have accumulated sequence differences that are most likely to result in a difference in the functional potential of a protein. It presents niche adaptation in bacteria as an evolutionary scenario that involves the loss of function of an array of protein coding genes, and showcases the ease of identification of these loss of function events using the software.

Chapter Three begins to scale up this approach, comparing many evolved *E. coli* lines to a common ancestor. The aim of the study is to identify genes that have accumulated deleterious mutations both under conditions of severe genetic bottlenecking producing strong genetic drift, and serial culture without severe bottlenecking, allowing greater genetic diversity in the population to improve the effectiveness of selection. The approach allows us to demonstrate that bacterial populations subjected to strong genetic drift accumulate more deleterious mutations than those that have a larger effective population size. This chapter also raises the question of how to distinguish adaptive, neutral and deleterious changes on the molecular level in recently diverged bacteria.

Chapter Four presents a further improvement to the scalability of the method, transitioning from comparing individual scores between isolates to comparing distributions of scores between bacteria occupying different niches. This is also the first study in the thesis to incorporate random forest methodology to identify a small number of informative genes in a sparse dataset. It is the first application of the method to clinical isolates of bacteria, comparing invasive *Campylobacter jejuni* isolates to reference gastrointestinal strains in order to identify genetic differences that may underlie adaptation to invasive infection.

Chapter Five presents an investigation of the evolution of invasive potential in *Salmonella enterica*, an extension of the work performed in Chapter Two, but with the new, group-based implementation of the method and a change of focus from host adaptation to invasiveness. The study takes the concept of p >> n near to its extreme, looking at a small set of carefully selected training data, allowing us to reconstruct the patterns that have been identified and explore the potential for overtraining and confounding due to population structure in larger datasets.

Chapter Six tests the method by taking a step further backward in time to look at functional differences between species, rather than strains. Specifically, we look at whether there are different degrees of sequence constraint on proteins found in pathogens versus nonpathogens in the genus *Pseudomonas*. There is strong confounding due to population structure in this chapter, causing the sequence changes associated with pathogenicity to be closely linked with lineage-specific changes. We argue that the nature of the delta-bitscore method allows us to detect the most functionally significant changes associated with a specific lineage, increasing the likelihood that sequence changes will relate to phenotype.

Chapter Seven examines a potential future avenue for synthesis of delta-bitscore (DBS) data with other large datasets - an investigation into differences in substrate utilisation efficiencies in different clinical isolates of *Staphylococcus* using Biolog phenotype arrays. It identifies some technical issues with processing Biolog data in large quantities and compares phylogenetic relationships between *Staphylococcus* isolates with their metabolic profiles. Importantly, it identifies measurable changes in phenotype occurring across short evolutionary distances that could be further probed using DBS and whole-genome sequencing.

Chapter Eight then summarises the biological and methodological findings of the thesis, and poses questions about the potential for expansion of this work, and the considerations that will need to be made to ensure investigations are rigorous and conclusions are valid.

References

Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Current Protocols in Human Genetics*, *Chapter 7*, Unit7.20.

Amaratunga, D., Cabrera, J., & Lee, Y.-S. (2008). Enriched random forests. *Bioinformatics* , *24*(18), 2010–2014.

Babyak, M. A. (2004). What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine*, *66*(3), 411–421.

Baker, M. G., Barnard, L. T., Kvalsvig, A., Verrall, A., Zhang, J., Keall, M., … Howden-Chapman, P. (2012). Increasing incidence of serious infectious diseases and inequalities in New Zealand: a national epidemiological study. *The Lancet*, *379*(9821), 1112–1119.

Barbosa, E., Röttger, R., Hauschild, A.-C., Azevedo, V., & Baumbach, J. (2014). On the limits of computational functional genomics for bacterial lifestyle prediction. *Briefings in Functional Genomics*, *13*(5), 398–408.

Breiman, L. (1996). *Out-of-bag estimation*. Statistics Department, University of California Berkeley, Berkeley CA 94708.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32.

Bureau, A., Dupuis, J., Falls, K., Lunetta, K. L., Hayward, B., Keith, T. P., & Van Eerdewegh, P. (2005). Identifying SNPs predictive of phenotype using random forests. *Genetic Epidemiology*, *28*(2), 171–182.

Caruana, R., Karampatziakis, N., & Yessenalina, A. (2008). An empirical evaluation of supervised learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning* (pp. 96–103). ACM.

Chen, L., Todd, R., Kiehlbauch, J., Walters, M., & Kallen, A. (2017). Notes from the Field: Pan-Resistant New Delhi Metallo-Beta-Lactamase-Producing Klebsiella pneumoniae - Washoe County, Nevada, 2016. *MMWR. Morbidity and Mortality Weekly Report*, *66*(1), 33.

Choi, Y., & Chan, A. P. (2015). PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* , *31*(16), 2745–2747.

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., & Chan, A. P. (2012). Predicting the functional effect of amino acid substitutions and indels. *PloS One*, *7*(10), e46688.
Christaki, E. (2015). New technologies in predicting, preventing and controlling emerging infectious diseases. *Virulence*, *6*(6), 558–565.
Clifford, R. J., Edmonson, M. N., Nguyen, C., & Buetow, K. H. (2004). Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. *Bioinformatics* , *20*(7), 1006–1014.
Dutilh, B. E., Thompson, C. C., Vicente, A. C. P., Marin, M. A., Lee, C., Silva, G. G. Z., … Edwards, R. A. (2014). Comparative genomics of 274 Vibrio cholerae genomes reveals mobile functions structuring three niche dimensions. *BMC Genomics*, *15*, 654.
Eddy, S. R. (1998). Profile Hidden Markov Models. *Bioinformatics*, *14*(9), 755–763.
Fauci, A. S., & Morens, D. M. (2012). The perpetual challenge of infectious diseases. *The New England Journal of Medicine*, *366*(5), 454–461.
Field, D., Wilson, G., & van der Gast, C. (2006). How do we compare hundreds of bacterial genomes? *Current Opinion in Microbiology*, *9*(5), 499–504.
Georgiades, K., & Raoult, D. (2011a). Defining Pathogenic Bacterial Species in the Genomic Era. *Frontiers in Microbiology*, *1*. https://doi.org/10.3389/fmicb.2010.00151
Georgiades, K., & Raoult, D. (2011b). Genomes of the most dangerous epidemic bacteria have a virulence repertoire characterized by fewer genes but more toxin-antitoxin modules. *PloS One*, *6*(3), e17962.
Gnad, F., Baucom, A., Mukhyala, K., Manning, G., & Zhang, Z. (2013). Assessment of computational methods for predicting the effects of missense mutations in human cancers. *BMC Genomics*, *14 Suppl 3*, S7.
Goldstein, B. A., Hubbard, A. E., Cutler, A., & Barcellos, L. F. (2010). An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings. *BMC Genetics*, *11*, 49.
Goodhead, I., & Darby, A. C. (2015). Taking the pseudo out of pseudogenes. *Current Opinion in Microbiology*, *23*, 102–109.
Gough, J., & Chothia, C. (2002). SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Research*, *30*(1), 268–272.
Haury, A.-C., Gestraud, P., & Vert, J.-P. (2011). The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PloS One*, *6*(12), e28210.
Hottes, A. K., Freddolino, P. L., Khare, A., Donnell, Z. N., Liu, J. C., & Tavazoie, S. (2013). Bacterial adaptation through loss of function. *PLoS Genetics*, *9*(7), e1003617.
Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*. https://doi.org/10.1177/2053951714528481
Koonin, E. V. (2016). Splendor and misery of adaptation, or the importance of neutral null for understanding evolution. *BMC Biology*, *14*(1), 114.
Kuo, C.-H., & Ochman, H. (2010). The extinction dynamics of bacterial pseudogenes. *PLoS Genetics*, *6*(8). https://doi.org/10.1371/journal.pgen.1001050
Law, J. W.-F., Ab Mutalib, N.-S., Chan, K.-G., & Lee, L.-H. (2014). Rapid methods for the detection of foodborne bacterial pathogens: principles, applications, advantages and limitations. *Frontiers in Microbiology*, *5*, 770.
Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). Big data. The parable of Google Flu: traps in big data analysis. *Science*, *343*(6176), 1203–1205.
Liao, J. G., & Chin, K.-V. (2007). Logistic regression for disease classification using microarray data: model selection in a large p and small n case. *Bioinformatics* , *23*(15), 1945–1951.
Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22.
Loh, W.-Y. (2012). Variable Selection for Classification and Regression in Large p, Small n Problems. In A. Barbour, H. P. Chan, & D. Siegmund (Eds.), *Probability Approximations and Beyond* (pp. 135–159). Springer New York.
Maurelli, A. T. (2007). Black holes, antivirulence genes, and gene inactivation in the evolution of bacterial pathogens. *FEMS Microbiology Letters*, *267*(1), 1–8.
Medina, M., & Sachs, J. L. (2010). Symbiont genomics, our new tangled bank. *Genomics*, *95*(3), 129–137.
Merhej, V., Georgiades, K., & Raoult, D. (2013). Postgenomic analysis of bacterial pathogens repertoire reveals genome reduction rather than virulence factors. *Briefings in Functional Genomics*, *12*(4), 291–304.
Merhej, V., Royer-Carenzi, M., Pontarotti, P., & Raoult, D. (2009). Massive comparative genomic analysis reveals convergent evolution of specialized bacteria. *Biology Direct*, *4*(1), 13.
Mira, A., Ochman, H., & Moran, N. A. (2001). Deletional bias and the evolution of bacterial genomes. *Trends in Genetics: TIG*, *17*(10), 589–596.
Morens, D. M., & Fauci, A. S. (2013). Emerging infectious diseases: threats to human health and global stability. *PLoS Pathogens*, *9*(7), e1003467.
Motsinger, A. A., & Ritchie, M. D. (2006). Multifactor dimensionality reduction: an analysis strategy for modelling and detecting gene-gene interactions in human genetics and pharmacogenomics studies. *Human Genomics*, *2*(5), 318–328.
Ng, P. C., & Henikoff, S. (2006). Predicting the effects of amino acid substitutions on protein function. *Annual Review of Genomics and Human Genetics*, *7*, 61–80.
Nuccio, S.-P., & Baumler, A. J. (2014). Comparative Analysis of Salmonella Genomes Identifies a Metabolic Network for Escalating Growth in the Inflamed Gut. *mBio*, *5*(2), e00929–14–e00929–14.
O'Neill, J. (2016). *Tackling drug-resistant infections globally: final report and recommendations*. London: Wellcome Trust & HM Government. Retrieved from

https://amr-review.org/sites/default/files/160525_Final%20paper_with%20cover.pdf
Pallen, M. J., & Wren, B. W. (2007). Bacterial pathogenomics. *Nature*, *449*(7164), 835–842.
Parkhill, J., Sebaihia, M., Preston, A., Murphy, L. D., Thomson, N., Harris, D. E., … Maskell, D. J. (2003). Comparative analysis of the genome sequences of Bordetella pertussis, Bordetella parapertussis and Bordetella bronchiseptica. *Nature Genetics*, *35*(1), 32–40.
Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., … Carroll, M. W. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature*, *530*(7589), 228–232.
Reva, B., Antipin, Y., & Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Research*, *39*(17), e118.
Rohmer, L., Fong, C., Abmayr, S., Wasnick, M., Larson Freeman, T. J., Radey, M., … Brittnacher, M. J. (2007). Comparison of Francisella tularensis genomes reveals evolutionary events associated with the emergence of human pathogenic strains. *Genome Biology*, *8*(6), R102.
Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L. A., Edwards, K. J., … Gaunt, T. R. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Human Mutation*, *34*(1), 57–65.
Siegel, E. (2013). *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, Or Die*. John Wiley & Sons.
Sjolander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, 1. Saira, & Haussler, D. (1995). Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Eddy*, *1995*, 1996.
Sonnhammer, E. L., Eddy, S. R., & Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, *28*(3), 405–420.
Tarca, A. L., Carey, V. J., Chen, X.-W., Romero, R., & Drăghici, S. (2007). Machine learning and its applications to biology. *PLoS Computational Biology*, *3*(6), e116.
Touw, W. G., Bayjanov, J. R., Overmars, L., Backus, L., Boekhorst, J., Wels, M., & van Hijum, S. A. F. T. (2013). Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? *Briefings in Bioinformatics*, *14*(3), 315–326.
van Nimwegen, K. J. M., van Soest, R. A., Veltman, J. A., Nelen, M. R., van der Wilt, G. J., Vissers, L. E. L. M., & Grutters, J. P. C. (2016). Is the $1000 Genome as Near as We Think? A Cost Analysis of Next-Generation Sequencing. *Clinical Chemistry*, *62*(11), 1458–1464.
Wheeler, N. E., Barquist, L., Kingsley, R. A., & Gardner, P. P. (2016). A profile-based method for identifying functional divergence of orthologous genes in bacterial genomes. *Bioinformatics* , *32*(23), 3566–3574.
World Health Organization. (2013). *World Health Report 2013: Research for Universal Health Coverage*. World Health Organization. Retrieved from http://apps.who.int/iris/bitstream/10665/85761/2/9789240690837_eng.pdf?ua=1
Wu, H.-J., Wang, A. H.-J., & Jennings, M. P. (2008). Discovery of virulence factors of pathogenic bacteria. *Current Opinion in Chemical Biology*, *12*(1), 93–101.
Yoon, Y., Song, J., Hong, S. H., & Kim, J. Q. (2003). Analysis of multiple single nucleotide polymorphisms of candidate genes related to coronary heart disease susceptibility by using support vector machines. *Clinical Chemistry and Laboratory Medicine: CCLM / FESCC*, *41*(4), 529–534.
Zhang, M., Zhang, D., & Wells, M. T. (2008). Variable selection for large p small n regression models with incomplete data: mapping QTL with epistases. *BMC Bioinformatics*, *9*, 251.

Chapter Two | A profile-based method for identifying functional divergence of orthologous genes in bacterial genomes

Preface

This chapter outlines the development and testing of the delta-bitscore (DBS) method of identifying mutations in protein-coding genes that are likely to result in changes to protein function. The method was first conceived by Lars Barquist and colleagues, and used to investigate genetic changes related to host adaptation in bacteria. The purpose of this thesis has been to test the performance of this approach and its competitiveness against other similar approaches. To do this, I collected benchmarking data consisting of results from several protein mutagenesis experiments. In performing this benchmarking I discovered that the original approach that employed models from the Pfam database performed poorly compared to other BLAST-based methods, so I developed a more competitive approach that uses custom-built profile HMMs. To showcase the utility of the method for comparative genomics applications, I used data from a study that had been published on the degradation of biochemical pathways in invasive *Salmonella* to demonstrate that our approach achieved similar results to a more labour-intensive, manual approach to analysis.

This chapter is published as follows:
Nicole E. Wheeler, Lars Barquist, Robert A. Kingsley, Paul P. Gardner; A profile-based method for identifying functional divergence of orthologous genes in bacterial genomes. *Bioinformatics* 2016; 32 (23): 3566-3574. doi: 10.1093/bioinformatics/btw518

The main and supplementary text are provided here, however additional supplementary material is available online at:
https://academic.oup.com/bioinformatics/article/32/23/3566/2525633/A-profile-based-method-for-identifying-functional#supplementary-data

Contributions

Lars Barquist wrote the software, I tested the software, performed the analyses and wrote the manuscript, all authors contributed to the design of the analysis.

Bioinformatics, 32(23), 2016, 3566–3574
doi: 10.1093/bioinformatics/btw518
Advance Access Publication Date: 8 August 2016
Original Paper

OXFORD

Sequence analysis

A profile-based method for identifying functional divergence of orthologous genes in bacterial genomes

Nicole E. Wheeler[1,2,†,*], Lars Barquist[3,†,*], Robert A. Kingsley[4,5] and Paul P. Gardner[1,2,6]

[1]School of Biological Sciences, University of Canterbury, Christchurch, New Zealand, [2]Biomolecular Interaction Centre, University of Canterbury, Christchurch, New Zealand, [3]Institute for Molecular Infection Biology, University of Wuerzburg, Wuerzburg, Germany, [4]Institute of Food Research, Norwich Research Park, Norwich, UK, [5]Wellcome Trust Sanger Institute, Hinxton, UK and [6]Bio-protection Research Centre, University of Canterbury, Christchurch, New Zealand

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate editor: Inanc Birol

Received on May 3, 2016; revised on July 17, 2016; accepted on August 2, 2016

Abstract

Motivation: Next generation sequencing technologies have provided us with a wealth of information on genetic variation, but predi
cting the functional significance of this variation is a difficult task. While many comparative genomics studies have focused on gene flux and large scale changes, relatively little attention has been paid to quantifying the effects of single nucleotide polymorphisms and indels on protein function, particularly in bacterial genomics.
Results: We present a hidden Markov model based approach we call delta-bitscore (DBS) for identifying orthologous proteins that have diverged at the amino acid sequence level in a way that is likely to impact biological function. We benchmark this approach with several widely used datasets and apply it to a proof-of-concept study of orthologous proteomes in an investigation of host adaptation in *Salmonella enterica*. We highlight the value of the method in identifying functional divergence of genes, and suggest that this tool may be a better approach than the commonly used dN/dS metric for identifying functionally significant genetic changes occurring in recently diverged organisms.
Availability and Implementation: A program implementing DBS for pairwise genome comparisons is freely available at: https://github.com/UCanCompBio/deltaBS.
Contact: nicole.wheeler@pg.canterbury.ac.nz or lars.barquist@uni-wuerzburg.de
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genome sequencing technologies allow us to explore the wealth of genetic variation between and within species, and as these technologies advance this data is becoming progressively cheaper, faster and easier to produce (Koren and Phillippy, 2015; Loman and Pallen, 2015; Loman *et al.*, 2012). However, analysis of the functional impact of genetic variation has lagged behind, and has largely focused on the presence or absence of macroscopic features such as particular genes, genomic islands or plasmids. Comparative sequence analyses have become common, and exploration of genetic variation between closely related organisms has provided key insights into bacterial evolution (Barquist and Vogel, 2015;

Bryant *et al.*, 2012; Croucher and Didelot, 2014). In particular, the analysis of single nucleotide polymorphisms (SNPs) has been a tremendous boon to the study of bacterial populations, allowing for the construction of phylogenetic trees which provide information on disease transmission and adaptation at scales ranging from global pandemics (Mutreja *et al.*, 2011) to outbreaks within single hospital wards (Harris *et al.*, 2013). Still, the functional analysis of these SNPs, insertions and deletions within protein sequences remains difficult, and often relies on inappropriate tools such as dN/dS.

How then can the significance of fine-scale genetic variation be quantified and prioritized for investigation? Recent studies have shown that even single SNPs can have dramatic effects on major phenotypes such as host tropism (Singletary *et al.*, 2016; Viana *et al.*, 2015; Yue *et al.*, 2015). Studies of pathogen adaptation, for example the adaptation of *Pseudomonas aeruginosa* to the cystic fibrosis lung (Jorth *et al.*, 2015; Marvig *et al.*, 2015) or of *Salmonella enterica* to immunocompromised populations (Feasey *et al.*, 2012; Okoro *et al.*, 2012, 2015), often result in findings of hundreds to thousands of SNPs and small indels in coding regions. Genome-wide association studies provide one method (Chewapreecha *et al.*, 2014) for interpreting this variation, however the clonal nature of many pathogens can make such study designs difficult if not impossible to pursue and require large sample sizes to be effective. The development of fast and accurate ways to assess the functional impact of variation between strains and prioritize coding variants for follow-up work is an important step in extracting meaning from comparative analyses.

Our strategy uses a profile HMM-based approach. Profile HMMs are probabilistic models of multiple sequence alignments. For each column in the alignment they capture information on the expected frequency of occurrence of different amino acids, insertions, and deletions. We can then use this information to compute a score, which we call delta-bit score (DBS) for reasons explained below, that quantifies the divergence of two protein sequences with respect to the conservation patterns captured by the profile HMM. Our key assumption in this is that variation in positions conserved across a protein family is more likely to affect protein function than variation in less conserved positions. Similar assumptions have had success in a number of applications such as the prediction of tRNA pseudogenes (Lowe and Eddy, 1997) and protein folds (Marks *et al.*, 2012). Our approach is unique in both the simplicity of the measure used, and its flexibility, making it well suited to comparative genomic applications.

We demonstrate that our method is effective using several commonly used protein mutagenesis datasets, particularly when used with automatically constructed protein alignments. We provide an example of a comparative genomic analysis, by applying our approach to a whole proteome analysis of a well-studied case of host adaptation in *Salmonella enterica* serovars Enteritidis and Gallinarum, using a purpose built collection of HMMs. We additionally show these results generalize to other host-adapted salmonellae. Finally, we compare our method to a popular estimate of selection, dN/dS, and find that the two measures are only weakly correlated. This adds to the evidence that dN/dS is not an appropriate measure for inferring selection for gene function within bacterial populations.

2 Materials and methods

2.1 Benchmarking

In order to test the ability of profile HMMs to identify loss-of-function mutations we used three independent datasets from protein mutagenesis studies on HIV-1 protease (336 sequences) (Loeb *et al.*, 1989), *Escherichia coli* Lac I (4041 sequences) (Markiewicz *et al.*, 1994) and phage lysozyme (2015 sequences) (Rennell *et al.*, 1991). In these experiments, residues in the protein were systematically mutated and the impact of these mutations on protein function was quantified. The data were downloaded from the SIFT website (http://sift.bii.a-star.edu.sg/) (Kumar *et al.*, 2009). For binary classification of mutants, proteins with a classification of '+' were termed functional mutants and '+−', '−+' and '—' were termed loss of function mutants. We tested two different reference HMM sets: curated HMMs from the Pfam database (Punta *et al.*, 2012), and automatically constructed HMMs built using a range of residue identity cutoff values. We also compared these results to predictions made by the PROVEAN, SIFT and Mutation Assessor methods. PROVEAN results were downloaded precomputed from their website. Otherwise, all methods were tested using their default settings. Exclusion of positions in the protein with low coverage in the SIFT database have only a marginal effect on SIFT performance (0.7–0.71 AUC in lysozyme dataset), and results in loss of data, so we use all sequence variants in our analysis. In order to evaluate the accuracy of each method we computed the 'Area Above the Curve' (AAC) for a series of Relative Cost Curve (RCC) plots (Montvida and Klawonn, 2014). These values range between one and zero with an AAC equal to one indicating a perfect prediction tool across the range of relative costs tested. The Biocomb package (www.cran.r-project.org/package=Biocomb) was used with default relative costs unless otherwise specified to calculate AAC scores, and the ROCR package (Sing *et al.*, 2005) was used to calculate other performance metrics, which can be found in Supplementary Table S1.

For the construction of custom HMMs, query sequences were searched against the Uniref90 database (Suzek *et al.*, 2015) using a single iteration of jackhmmer (Eddy, 2011). Pairwise sequence identity was calculated as the number of matches between sequence pairs, divided by total length of the alignment after the removal of gap-gap columns. HMMs were built using hmmbuild, and DBS was calculated using hmmsearch, subtracting the bitscore value for the variant sequence from the bitscore value for the wild type sequence. jackhmmer, hmmbuild and hmmsearch are part of the HMMER3 package (version 3.1b2). Protein HMMs were built using a range of sequence identity cutoffs in order to determine the optimal range for performance.

2.2 Proteome analysis

We have designed this approach with whole proteome analyses in mind, so as a proof-of-concept, we applied the approach to the study of host adaptation in *Salmonella enterica*, using *S.* Enteritidis and *S.* Gallinarum as a specific example. For a summary of the workflow of this portion of the analysis, see Supplementary Figure 1. Genomes for *Salmonella* Enteritidis str. P125109 (AM933172.1) and Salmonella Gallinarum str. 287/91 (AM933173.1) were retrieved. Custom gene models were constructed by searching each *S.* Enteritidis protein coding gene against the Uniref90 database. Profiles were built using sequences showing 40% or greater sequence identity to the original query protein.

In order to assess whether models built from a small number of sequences performed well, we tested the performance of models using proteins from the humsavar database (http://www.uniprot.org/docs/humsavar), which catalogs human polymorphisms and disease variants for a wide range of proteins which have different levels of representation in Uniprot90, and therefore result in profile models built from varying numbers of sequences. We built custom

models for each protein in the database, then separated the models into classes based on both number of sequences and effective sequence number. We then computed AUC values for predictions made on variant data from each class. Results from this test can be found in Supplementary Table S2. We saw no classes where custom models performed worse than Pfam models, so we did not filter models built from few sequences. We did, however, use Pfam models for those proteins with no homologs in Uniref90 rather than build a model from the query sequence, as we felt this could bias results towards the reference species.

To assess the quality of our analysis, we wished to compare our results to those of Nuccio and Bäumler (2014), so we used the ortholog calls provided in their supplementary material. Genes represented by a custom model were scored using the appropriate model ($n = 3154$), all others were scored using Pfam domain models. hmmsearch was used for scoring. If hits to multiple Pfam models occurred, any overlapping hits were competed based on *E*-value. Orthologs with incompatible Pfam domain architectures ($n = 32$) were excluded from scoring, but counted as hypothetically attenuated coding sequences (HACs) in Figure 2A if they involved a loss of domain in one serovar. We anticipate that the variance of delta-bitscores will increase with evolutionary distance, so rather than establish a fixed scoring cutoff for HACs, we identify loss-of-function mutations using an empirical distribution. We set a score threshold at which 2.5% of genes on the least dispersed side of the distribution would be called as HACs. If the two bacteria show an equal rate of protein function loss, this would result in 5% of orthologous proteins being called as HACs, however if one proteome shows a greater degree of functional degradation, this will result in a greater proportion of orthologs being classified as HACs.

Functional classification was performed using pathways from the KEGG database (Kanehisa *et al.*, 2016). We grouped genes into four categories: those present in a pathway but with no ortholog in the other serovar; genes identified as hypothetically disrupted coding sequences (HDCs) by Nuccio and Bäumler (2014); genes identified by our DBS method as HACs, but not as HDCs; and finally genes not identified as non-functional by either method. dN/dS values were calculated using PAML (Yang, 1997), and for the comparison of DBS and dN/dS, genes were filtered for those with $dN > 0$ and $dS > 0.0001$. Correlations between measures were computed using a Spearman's rho statistic (R package cor).

In our investigation of multiple *S. enterica* isolates, for all orthologous groups with a gene present in *S.* Enteritidis, scores were collated, and if individual scores were significantly different to the median score for all isolates, we identified these proteins as HACs. Significant difference was calculated in a similar way to the pairwise comparisons, with the score corresponding to the most extreme 2.5% of delta-bitscores on the least dispersed side of an empirical distribution being used as the cutoff.

3 Results

3.1 The delta-bitscore measure

Alignment of a protein sequence to a profile HMM produces a bitscore value, which is the log odds ratio for this sequence under the profile HMM compared to a random null model, and serves as an indication of protein family membership. In subtracting the bitscore of one protein or domain from that of another, we produce a measure of the divergence between the two proteins, taking into consideration conservation patterns captured by the protein family model.

We define delta-bitscore using the following equation:

$$\mathrm{DBS} = \mathrm{x}_{\mathrm{ref}} - \mathrm{x}_{\mathrm{var}}$$

where DBS is delta-bitscore and, $\mathrm{x}_{\mathrm{ref}}$ and $\mathrm{x}_{\mathrm{var}}$ are bitscores for reference and variant sequences derived from alignments to the same profile HMM.

Highly conserved positions in a model alignment contribute more to bitscores than poorly conserved positions, meaning that unexpected mutations or indels in conserved sites are given a greater penalty than mutations in variable sites. In addition, the replacement of residues with chemically and structurally similar residues is generally scored more favorably than mutation to dissimilar residues. Multiple functionally neutral changes are likely to result in individual contributions to the bitscore that are small in magnitude and that cancel out over the length of the protein, while functionally significant change in a protein will likely produce one or more position-specific values of high magnitude that have a greater impact on overall DBS for the sequence. We first introduced this measure as part of studies of *Salmonella* adaptation to an wild avian host (Kingsley *et al.*, 2013), and during within-host evolution of a hypermutator strain of *Salmonella* Enteritidis in an immunocompromised patient (Klemm *et al.*, 2016), and elaborate upon it here. The Supplementary Text contains further discussion of this measure and comparisons with related measures derived from profile HMMs (Clifford *et al.*, 2004; Liu *et al.*, 2015; Shihab *et al.*, 2013).

3.2 DBS is predictive of protein functional status

Evaluating the performance of different methods for the quantification of the effects of sequence variation is challenging. For the case of human variation, some datasets exist, however given the limited amount of data available on functional protein variants care must be taken to avoid circularity in training and testing (Boulesteix, 2010; Grimm *et al.*, 2015). While we have not comprehensively benchmarked DBS on human data sets given our focus on prokaryotic variation, we found it is competitive with other untrained methods (Supplementary Figs S3 and S4), despite its simplicity. In the case of bacterial variation, we are not aware of well-characterized collections of protein variants. Rather, to test our method, we compared its performance to a selection of methods on three large scale protein mutagenesis datasets (Kumar *et al.*, 2009) (Fig. 1A; Supplementary Fig. S2 shows AUC values for the same analysis).

A number of methods have been designed for predicting the impact of sequence variation on the function of non-human proteins, and we compare our measure to three competitive methods here (Grimm *et al.*, 2015). These include: PROVEAN, which uses a BLAST-based approach to score sequence variants using closely related sequences (Choi *et al.*, 2012); the SIFT algorithm which uses position-specific scoring matrices based on sequence homology and known patterns of common amino acid substitutions to predict the functional consequences of non-synonymous single nucleotide polymorphisms (nsSNPs) (Kumar *et al.*, 2009); and MutationAssessor which computes both the conservation and specificity of residues within protein subfamilies to assess the impact of a mutation (Reva *et al.*, 2011).

As a comparison, we used DBS in two distinct ways. In the first, we calculated DBS scores based on alignments to curated Pfam domains (Punta *et al.*, 2012). In the second, we constructed profile-HMMs containing sequences of varying minimal sequence identity. These two approaches to model construction capture distinct forms of information: Pfam domains are designed to capture maximal

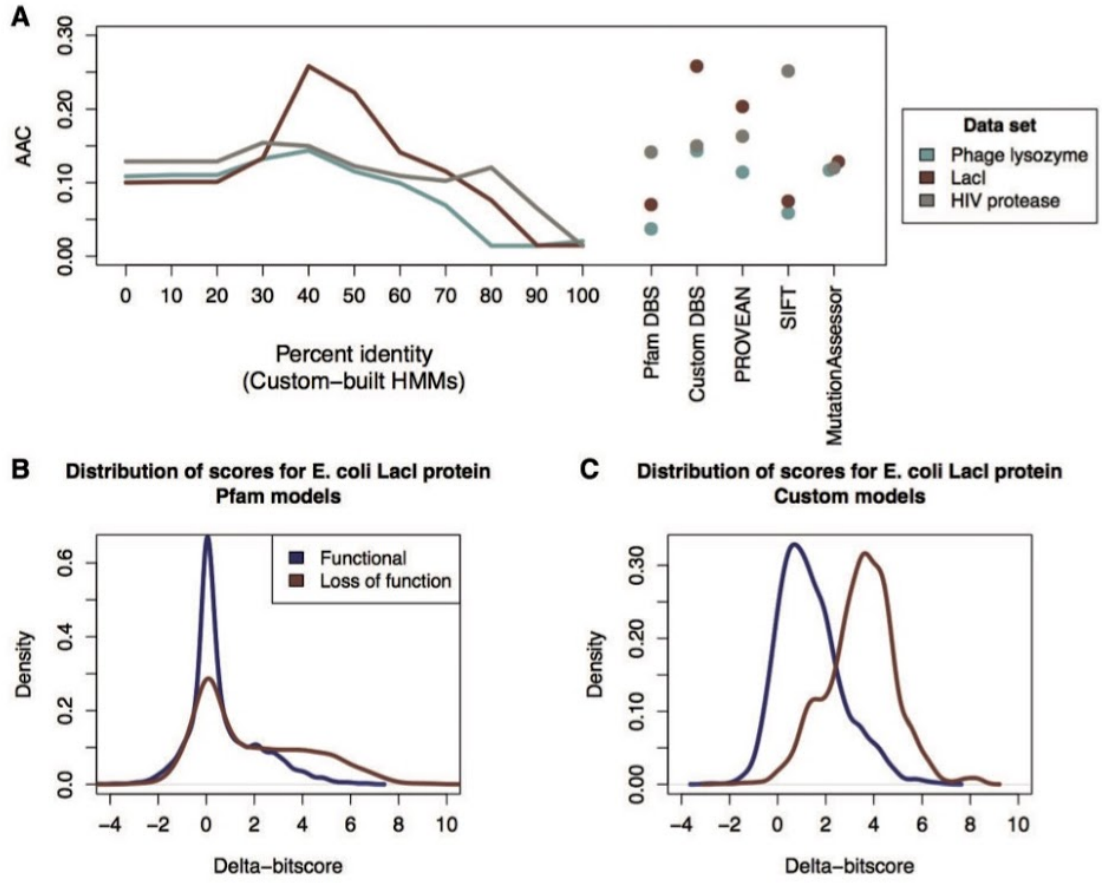


Fig. 1. (A) Area above the relative cost curve (AAC) values for different analysis methods using three mutagenesis benchmarking datasets: phage lysozyme, LacI and HIV protease. The performance for each model across a range of sequence identity cutoffs is shown, as well as AAC values for the individual methods. (B) Distribution of delta-bitscores for *E. coli* LacI using custom model analysis; (C) Distribution of delta-bitscores for *E. coli* LacI using Pfam model analysis

sequence diversity, and focus on deeply conserved domains within protein sequences. Additionally, many families are manually curated, and the alignment of known functional regions, such as active sites, may have been improved in some cases. Our custom profiles capture full length protein sequences. This means that the functions of sequences captured by custom models may be narrower than those captured by Pfam, and additionally will capture linker regions between conserved domains, which may be functionally important.

Our benchmark showed that custom HMMs built at 40% identity outperformed other methods in terms of area above the relative cost curve (AAC) and maximum Matthew's correlation coefficient, with the exception of the superior performance of SIFT and PROVEAN on the HIV protease data (Fig. 1A and Supplementary Table S1). DBS using Pfam models had the best specificity across all datasets tested, but at the expense of sensitivity (see Fig. 1B and Supplementary Table S1). This suggests that Pfam models are indeed capturing ancient features of protein families, and may be useful in cases where a low false-positive rate is desired at the expense of potentially missing variants. Our custom models meanwhile appeared to perform optimally with a 40% identity cutoff (Fig. 1A and C). Interestingly, 40% sequence identity has previously been proposed as a rule of thumb identity cutoff for the transfer of enzyme annotations (Addou *et al.*, 2009; Tian and Skolnick, 2003), indicating that this cutoff may reflect a common point of functional divergence in protein families.

3.3 DBS identifies pathways associated with host range restriction in *Salmonella* Gallinarum

We have established that DBS can detect functionally relevant mutations. To explore the utility of this approach in a comparative genomics context, we developed a tool that takes whole proteome files as input, builds custom HMMs where applicable and uses Pfam HMMs to score the remaining genes, in order to identify functionally significant variation between two proteomes. Using this tool, we compared the proteomes of two closely related *Salmonella enterica* serovars that have experienced contrasting selective pressures leading to host-restriction for one of them.

Host-restriction is a common phenomenon in highly adapted invasive pathogens, often characterized by genomic features such as the proliferation of transposable elements and the degradation of substantial fractions of coding sequences (Goodhead and Darby, 2015; Moran and Plague, 2004). Within *Salmonella enterica* such restriction events have occurred independently multiple times in various hosts from broad host-range ancestors. *Salmonella enterica* serovar Enteritidis is a broad host-range pathogen, capable of infecting humans, cattle, rodents and a variety of birds, while the

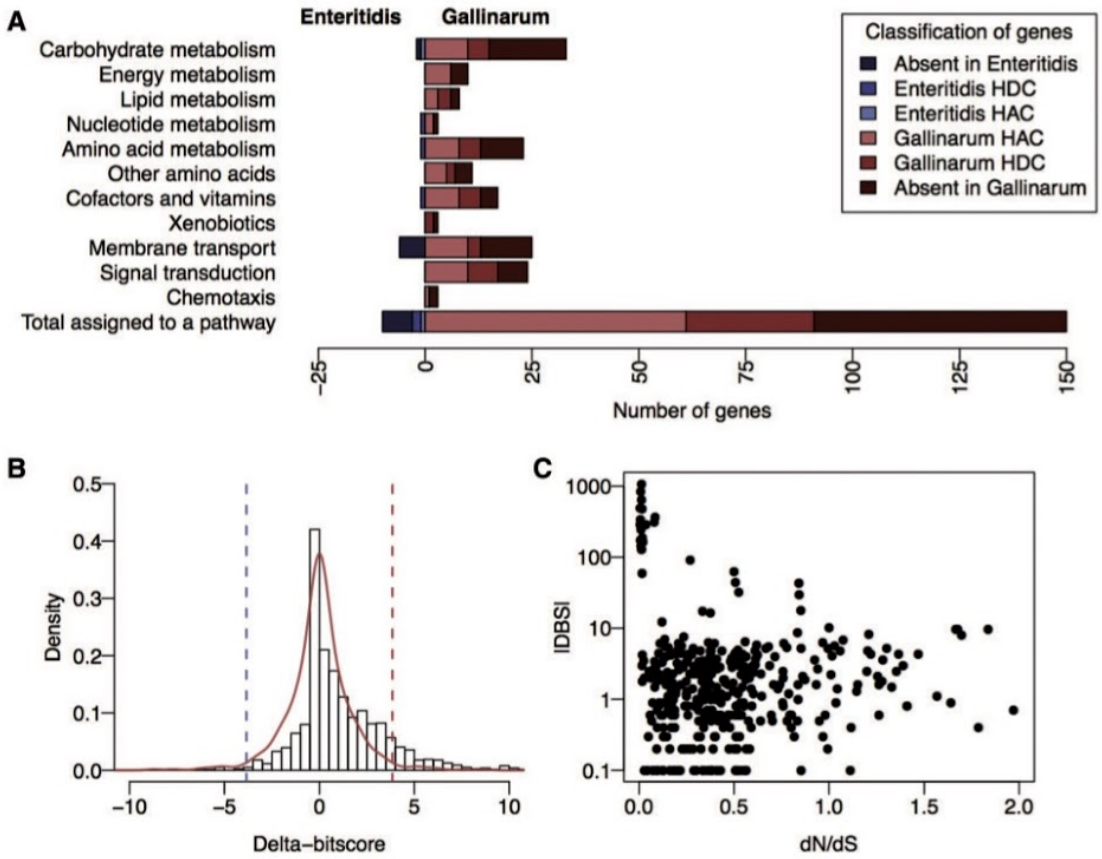


Fig. 2. The results of a DBS comparison of the proteomes of *S.* Enteritidis (a generalist infectious agent) and *S.* Gallinarum (a host-restricted infectious agent). (A) The functional changes in orthologous protein-coding genes of *S.* Enteritidis and *S.* Gallinarum have been grouped into functional categories using the KEGG pathways database. Genes included in the pathway that have no ortholog in the other serovar are indicated in the darkest colour, followed by genes previously identified as hypothetically disrupted coding sequences (HDCs), then genes with significant DBS values that had not already been identified as HDCs. We refer to these as hypothetically attenuated coding sequences (HACs); (B) The distribution of delta-bitscores for orthologous genes showing non-synonymous changes in *S.* Enteritidis and *S.* Gallinarum. A symmetrical empirical distribution of scores generated by mirroring the least dispersed side is shown in red, the cutoff values we use to establish significance are shown as dashed lines. A skewed distribution implies excess functional changes in one lineage; (C) A plot of the dN/dS score and corresponding DBS is shown for each orthologous *S.* Enteritidis and *S.* Gallinarum gene

closely related serovar Gallinarum is restricted to infecting galliforme birds (Rabsch *et al.*, 2002). *S.* Gallinarum and *S.* Enteritidis have recently evolved from a common ancestor, however the *S.* Gallinarum genome has undergone extensive degradation since divergence (Langridge *et al.*, 2015; Thomson *et al.*, 2008). In addition to being restricted to a narrow host range, *S.* Gallinarum has lost motility and causes a systemic, typhoid-like infection in birds, unlike Enteritidis which usually causes a self-limiting gastroenteritis (Thomson *et al.*, 2008). A recent analysis of pseudogenes within this and other invasive *Salmonella* lineages identified signatures of host-restriction, characterized particularly by the loss of metabolic genes required for survival in the intestine (Langridge *et al.*, 2015; Nuccio and Bäumler, 2014). Both of these analyses involved the manual comparison of coding sequences to identify mutations that would result in frameshifts or truncated proteins. We expect DBS to add an additional layer of information to such an analysis, identifying genes which have shifted in or lost function due to non-synonymous mutations or small indels occurring since the restriction event, but have not yet succumbed to obvious disruption events such as large truncations, frameshifts or complete deletions.

As shown in Figure 2A, our loss-of-function predictions included many genes not identified as hypothetically disrupted coding sequences (HDCs) by manual inspection (Nuccio and Bäumler, 2014).

We examined the concurrence between our loss of function calls and those made by Nuccio and Bäumler. Of 252 HDCs identified in *S.* Gallinarum but not in *S.* Enteritidis from the previous study, 148 of these were also classified as HACs by DBS, and a further 6 as HACs due to a loss of a domain in the *S.* Gallinarum copy of the gene (61% total agreement). Of the 104 disagreements, 28 genes had a DBS score below the threshold, and 70 were excluded from scoring analysis for having incompatible domain architectures ($n = 14$) or no hits for either protein sequence to the Pfam database or any of our custom models ($n = 56$). Of these 56, over half were fewer than 150 amino acids long, often due to extreme truncations or early frameshifts (see Supplementary Table S3). Of the 28 low scoring variants, many involved short truncations,

alternate starts or indels of <10 amino acids. The remainder involved *S.* Gallinarum pseudogenes that had no hits to custom models or Pfam domains in the truncated region. In many cases, our models showed low sequence conservation and gaps in the regions affected by mutation.

As expected, the distribution of DBS values centers around zero (Fig. 2B), showing that most orthologous gene pairs differ from the modeled sequence constraints to a similar degree. The distribution of DBS values shows an enrichment for positive DBS (exact binomial test, $P = 2.18\ e{-}35$), indicating greater divergence from the profile models for protein coding genes from *S.* Gallinarum when compared to *S.* Enteritidis. While some of these positive DBS values may indicate functional divergence rather than pseudogenization, extreme DBS values predominantly correspond to truncations, indels and mutations in highly conserved sites, so we assume that in these cases the ancestral function has been lost. To look at these hypothetically attenuated coding sequences (HACs) in a functional context, we grouped genes into functional categories based on their annotation in the KEGG database. Not only does *S.* Gallinarum have fewer genes than *S.* Enteritidis for most of the functional categories we considered (data not shown), but it also has a greater number of HACs across these groupings (Fig. 2A). Previous work that found the presence of non-ancestral pseudogenes in *S.* Enteritidis was limited while *S.* Gallinarum had accumulated a large number of pseudogenes since divergence is consistent with our results (Langridge *et al.*, 2015).

In our analysis, we found a number of results which are in agreement with previously published findings. Degradation of genes involved in chemotaxis is consistent with a loss of motility in this strain, and previously identified degradation of chemotaxis genes in other host adapted strains (Kingsley *et al.*, 2013; McClelland *et al.*, 2004). Fimbriae are also thought to be important for host colonization by *Salmonella*, and we found high DBS values for *bcfA*, *bcfC* and *stfG*. Pseudogenization of other genes in these operons have already been identified in *S.* Gallinarum (Foley *et al.*, 2013). We found a number of HACs in genes involved in the utilization of nutrients derived from the inflamed host gut environment colonized by gastrointestinal serovars of *Salmonella* (Nuccio and Bäumler, 2014) (Supplementary Table S5). This is consistent with the recent adaptation of *S.* Gallinarum from an ancestral gastrointestinal pathogen to an extra-intestinal environment. Among the most highly ranked genes were representatives of the *cbi*, *pdu* and *eut* operons, previously identified as central pathways subject to gene decay in *S.* Gallinarum (Langridge *et al.*, 2015; Nuccio and Bäumler, 2014). A complete table of DBS values for orthologous genes can be found in Supplementary Table S6.

These findings demonstrate that DBS provides an additional layer of information about gene function in addition to gene deletion when investigating serovars that have recently diverged. As time since divergence increases we expect that non-functional genes will be deleted and the ratio of non-functional and deleted genes will change. We anticipate that our method will be most useful in comparisons of organisms which have recently diverged, as it offers an opportunity to identify loss-of-function mutations that occur as an immediate response to a new environment and restricted population size, before deletion of entire genes occurs (Kuo and Ochman, 2010).

3.4 DBS does not correlate strongly with dN/dS

In comparing genome sequences, we have focused on identifying functionally significant variation, whereas a common approach is to classify genes from an evolutionary point of view, i.e. to identify genes are under negative selection, positive selection or that are evolving neutrally. A commonly used measure of selection is dN/dS (ω), which compares the rate of non-synonymous changes in protein-coding sequences to the rate of synonymous changes to classify genes as either negative selection ($\omega < 1$), positive selection ($\omega > 1$) or neutrally evolving ($\omega \approx 1$) (Yang and Bielawski, 2000). While dN/dS is an inappropriate metric to use in comparing bacteria at the strain level (Rocha *et al.*, 2006), the metric is nevertheless used in similar situations (Fleischmann *et al.*, 2002; Holden *et al.*, 2004; Roumagnac *et al.*, 2006), so we compared the results of the commonly used PAML program and DBS. In comparing the results from the two approaches, one of the most striking results is that there is little correlation between DBS and both dN/dS ($R \approx 0.17$) and dN ($R \approx 0.18$). Of the genes with high DBS and low dN/dS, most of these can be attributed to insertions and deletions, which are ignored by dN/dS but can have significant impacts on the functioning of a protein. Genes with low DBS and high dN/dS tend to carry a small number of non-synonymous changes, most of which are between chemically similar residues, and a smaller number of synonymous mutations.

3.5 DBS reveals trends of gene degradation across host-adapted *Salmonella*

To test whether a positively skewed DBS distribution was a common feature of host adaptation in *Salmonella*, we performed DBS analysis on three broad host range and three host-adapted *Salmonella* serovars. Rather than performing many pairwise comparisons to call HACs, for this analysis, we compared bitscores for each gene to the median gene bitscore of the six strains. We found that host-restricted serovars showed a greater number of HACs than generalist serovars across all strains tested (Fig. 3A). We previously observed a similar phenomenon in host-restricted strains of *S.* Typhimurium (Kingsley *et al.*, 2013). We observed a striking similarity in the score distributions of generalist strains, and in host-adapted strains (Supplementary Fig. S6), with the generalist scores being lower overall than the host restricted scores. It is interesting to note that most of

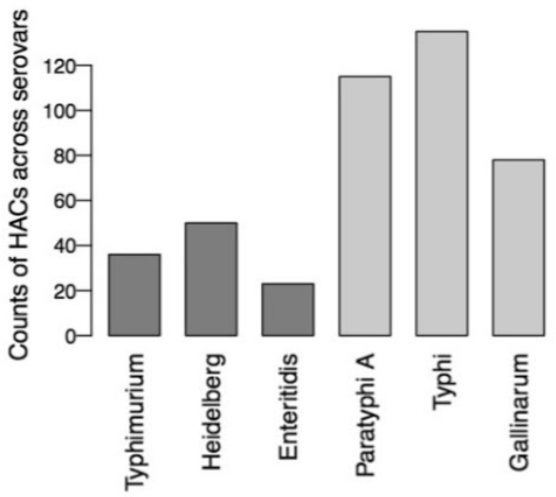


Fig. 3. Counts of hypothetically attenuated coding sequences (HACs) for each serovar, identified as the most extreme deviations from the median bitscore across strains for each gene. The generalist serovars are coloured in dark grey, host-restricted serovars are coloured in light grey

the separation in scores occurs at DBS values below our cutoff value for HAC classification (DBS = 8.2, Supplementary Fig. S6, top panel). Based on our benchmarking, if these proteins were tested *in vitro* many of them would show differences in function, but whether these differences relate to differences in intracellular and environmental conditions shaping the sequence requirements for these proteins or whether they relate to true declines in function is difficult to assess without experimental validation.

4 Discussion

We have described a surprisingly simple method for identifying functional divergence in orthologous proteins, including analysis of both point mutations and indel events. Despite the simplicity of the DBS measure, its accuracy is similar to state-of-the-art methods for predicting the functional impact of sequence variants. While curated Pfam models were less sensitive, they have a high specificity at conservative scoring thresholds (Supplementary Table S1 and Supplementary Fig. S5). The workflow for using such models is simpler (Supplementary Fig. S1), and may be more appropriate for investigations where a low false-positive rate is more important than detection of all loss-of-function mutations. In addition, while construction of custom models can be time-consuming to run on a desktop computer, Pfam annotation can be performed rapidly. Because other methods for estimating the impact of sequence variation, such as SIFT, PROVEAN and MutationAssessor, require the user to specify mutations individually, broad scale comparative proteome analysis such as this is difficult and time-intensive to perform, and not all methods are able to score indels or more than one mutation in the same sequence. Both DBS approaches scale to whole proteome analysis with minimal user involvement (see Github page for pipelines), and were able to detect additional loss of function mutations that were missed by previous studies focused on protein truncations and frameshifts.

4.1 DBS for the study of bacterial adaptation

Large scale genomic studies of bacterial pathogens have revealed striking similarities in evolutionary patterns in diverse pathogenic and symbiotic lineages. While adaptation to a new environment may involve the acquisition of new fitness determinants and rare, beneficial mutations, niche adaptation is also frequently accompanied by widespread loss-of-function mutations, particularly in pathogenic and symbiotic lineages. These loss-of-function mutations are likely generated through a number of distinct processes (Moran, 2002), including neutral or slightly deleterious stochastic loss due to reductions in effective population size, neutral loss of genes no longer required in the new environment, and adaptive loss (Hottes *et al.*, 2013). DBS presents an opportunity to mine the genomes of bacteria adapting to new environments for these loss of function mutations, which may tell us as much about bacterial adaptation to new environments as the acquisition of new genes. Comparative approaches based on pseudogene analysis have identified consistent signatures of adaptation in a number of bacterial pathogens, including *Yersinia* (McNally *et al.*, 2016; Reuter *et al.*, 2014), *E. coli* and *Shigella* (Feng *et al.*, 2011; Monk *et al.*, 2013) and *Salmonella* (Langridge *et al.*, 2015; Nuccio and Bäumler, 2014). Our analysis of *Salmonella* genomes has shown that DBS is consistent with these previous studies, while providing additional sensitivity to detect non-functional or functionally divergent protein variants that may have been missed by pseudogene analyses (Kingsley *et al.*, 2013).

4.2 DBS as a more appropriate alternative to dN/dS for studying evolution across short time scales

While dN/dS can be a powerful approach to identifying genes under positive selection across long time scales, on shorter timescales it has been shown to be inaccurate, due to the lag in the removal of slightly deleterious mutations This leads to high dN/dS ratios being commonplace in comparisons of closely related strains, suggesting positive or relaxed selection where there is none (Rocha *et al.*, 2006). Other studies have shown that dN/dS can provide inaccurate, even contradictory, results depending on the structure of the population being sampled (Kryazhimskiy and Plotkin, 2008). In spite of this, dN/dS is still commonly used in comparisons of closely related bacteria. Due to the unreliability of dN/dS measures at short evolutionary timescales and the inability of dN/dS based methods to score indels, we propose that DBS is a more suitable analysis tool for the study of functional divergence in recently diverged strains. In contrast, we would advise caution in the use of DBS on more divergent organisms, with careful consideration of the effects of sequence divergence on bitscore. Over time each lineage could accumulate mutations that affect protein function in different ways, but have the same degree of impact on bitscore, leading to a net DBS near zero. This does not necessarily indicate preservation of function and may instead indicate equal divergence from ancestral function. We have not examined the performance of DBS on deeply diverged lineages, partly for lack of data. Our observation that DBS performs optimally when reference alignments are in the 40% identity range suggests sequences more divergent than this may be outside the range where can DBS provide meaningful information, but this will require additional investigation to establish.

5 Concluding remark

We anticipate that this approach will be a useful addition to traditional comparative genomics workflows. It provides an effective method for scoring the functional potential of the overwhelming numbers of non-synonymous variants that can distinguish closely related bacteria, allowing better prioritization for further experimental investigation.

Acknowledgements

We would like to thank Fatemeh Ashari Ghomi for her contributions, and Sean Eddy for his clarification of some technical aspects of the HMMER3 software.

Funding

This work was supported in part by the Wellcome Trust, grant number WT098051. NEW is supported by a PhD scholarship from the University of Canterbury. LB was supported by a Research Fellowship from the Alexander von Humboldt Stiftung/Foundation. PPG and NEW are supported by a Rutherford Discovery Fellowship administered by the Royal Society of New Zealand.

Conflict of Interest: none declared.

References

Addou,S. *et al.* (2009) Domain-based and family-specific sequence identity thresholds increase the levels of reliable protein function transfer. *J. Mol. Biol.*, **387**, 416–430.

Barquist,L. and Vogel,J. (2015) Accelerating discovery and functional analysis of small RNAs with new technologies. *Annu. Rev. Genet.*, **49**, 367–394.

Boulesteix,A.L. (2010) Over-optimism in bioinformatics research. *Bioinformatics*, **26**, 437–439.

Bryant,J. *et al.* (2012) Developing insights into the mechanisms of evolution of bacterial pathogens from whole-genome sequences. *Future Microbiol.*, **7**, 1283–1296.

Chewapreecha,C. *et al.* (2014) Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes. *PLoS Genet.*, **10**, e1004547.

Choi,Y. *et al.* (2012) Predicting the functional effect of amino acid substitutions and indels. *PLoS One*, **7**, e46688.

Clifford,R.J. *et al.* (2004) Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. *Bioinformatics*, **20**, 1006–1014.

Croucher,N.J. and Didelot,X. (2014) The application of genomics to tracing bacterial pathogen transmission. *Curr. Opin. Microbiol.*, **23C**, 62–67.

Eddy,S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.

Feasey,N.A. *et al.* (2012) Invasive non-typhoidal salmonella disease: an emerging and neglected tropical disease in Africa. *Lancet*, **379**, 2489–2499.

Feng,Y. *et al.* (2011) Gene decay in Shigella as an incipient stage of host-adaptation. *PLoS One*, **6**, e27754.

Fleischmann,R.D. *et al.* (2002) Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. *J. Bacteriol.*, **184**, 5479–5490.

Foley,S.L. *et al.* (2013) Salmonella pathogenicity and host adaptation in chicken-associated serovars. *Microbiol. Mol. Biol. Rev.*, **77**, 582–607.

Goodhead,I. and Darby,A.C. (2015) Taking the pseudo out of pseudogenes. *Curr. Opin. Microbiol.*, **23C**, 102–109.

Grimm,D.G. *et al.* (2015) The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum. Mutat.*, **36**, 513–523.

Harris,S.R. *et al.* (2013) Whole-genome sequencing for analysis of an outbreak of meticillin-resistant *Staphylococcus aureus*: a descriptive study. *Lancet Infect. Dis.*, **13**, 130–136.

Holden,M.T.G. *et al.* (2004) Complete genomes of two clinical Staphylococcus aureus strains: evidence for the rapid evolution of virulence and drug resistance. *Proc. Natl. Acad. Sci. USA*, **101**, 9786–9791.

Hottes,A.K. *et al.* (2013) Bacterial adaptation through loss of function. *PLoS Genet.*, **9**, e1003617.

Jorth,P. *et al.* (2015) Regional isolation drives bacterial diversification within cystic fibrosis lungs. *Cell Host Microbe*, **18**, 307–319.

Kanehisa,M. *et al.* (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–D462.

Kingsley,R.A. *et al.* (2013) Genome and transcriptome adaptation accompanying emergence of the definitive type 2 host-restricted *Salmonella enterica* serovar Typhimurium pathovar. *MBio*, **4**, 13–e00565.

Klemm,E.J. *et al.* (2016) Emergence of host-adapted *Salmonella enteritidis* through rapid evolution in an immunocompromised host. *Nat. Microbiol.*, **1**, 15023.

Koren,S. and Phillippy,A.M. (2015) One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Curr. Opin. Microbiol.*, **23C**, 110–120.

Kryazhimskiy,S. and Plotkin,J.B. (2008) The population genetics of dN/dS. *PLoS Genet.*, **4**, e1000304.

Kumar,P. *et al.* (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.*, **4**, 1073–1081.

Kuo,C.H. and Ochman,H. (2010) The extinction dynamics of bacterial pseudogenes. *PLoS Genet.*, **6**(8): e1001050. doi:10.1371/journal.pgen.1001050

Langridge,G.C. *et al.* (2015) Patterns of genome evolution that have accompanied host adaptation in Salmonella. *Proc. Natl. Acad. Sci. USA*, **112**, 863–868.

Liu,M. *et al.* (2015) HMMvar-func: a new method for predicting the functional outcome of genetic variants. *BMC Bioinformatics*, **16**, 351.

Loeb,D.D. *et al.* (1989) Complete mutagenesis of the HIV-1 protease. *Nature*, **340**, 397–400.

Loman,N.J. *et al.* (2012) High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nat. Rev. Microbiol.*, **10**, 599–606.

Loman,N.J. and Pallen,M.J. (2015) Twenty years of bacterial genome sequencing. *Nat. Rev. Microbiol.*, **13**, 787–794.

Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.

Markiewicz,P. *et al.* (1994) Genetic studies of the lac repressor. XIV. Analysis of 4000 altered *Escherichia coli* lac repressors reveals essential and non-essential residues, as well as 'spacers' which do not require a specific sequence. *J. Mol. Biol.*, **240**, 421–433.

Marks,D.S. *et al.* (2012) Protein structure prediction from sequence variation. *Nat. Biotechnol.*, **30**, 1072–1080.

Marvig,R.L. *et al.* (2015) Convergent evolution and adaptation of *Pseudomonas aeruginosa* within patients with cystic fibrosis. *Nat. Genet.*, **47**, 57–64.

McClelland,M. *et al.* (2004) Comparison of genome degradation in Paratyphi A and Typhi, human-restricted serovars of *Salmonella enterica* that cause typhoid. *Nat. Genet.*, **36**, 1268–1274.

McNally,A. *et al.* (2016) 'Add, stir and reduce': Yersinia spp. as model bacteria for pathogen evolution. *Nat. Rev. Microbiol.*, **14**, 177–190.

Monk,J.M. *et al.* (2013) Genome-scale metabolic reconstructions of multiple *Escherichia coli* strains highlight strain-specific adaptations to nutritional environments. *Proc. Natl. Acad. Sci. USA*, **110**, 20338–20343.

Montvida,O. and Klawonn,K. (2014) Relative cost curves: an alternative to AUC and an extension to 3-class problems. *Kybernetika*, **50**, 647–660.

Moran,N.A. (2002) Microbial minimalism: genome reduction in bacterial pathogens. *Cell*, **108**, 583–586.

Moran,N.A. and Plague,G.R. (2004) Genomic changes following host restriction in bacteria. *Curr. Opin. Genet. Dev.*, **14**, 627–633.

Mutreja,A. *et al.* (2011) Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature*, **477**, 462–465.

Nuccio,S.P. and Bäumler,A.J. (2014) Comparative analysis of Salmonella genomes identifies a metabolic network for escalating growth in the inflamed gut. *MBio*, **5**, e00929–e00914.

Okoro,C.K. *et al.* (2012) Intracontinental spread of human invasive *Salmonella typhimurium* pathovariants in sub-Saharan Africa. *Nat. Genet.*, **44**, 1215–1221.

Okoro,C.K. *et al.* (2015) Signatures of adaptation in human invasive *Salmonella typhimurium* ST313 populations from sub-Saharan Africa. *PLoS Negl. Trop. Dis.*, **9**, e0003611.

Punta,M. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.

Rabsch,W. *et al.* (2002) *Salmonella enterica* serotype Typhimurium and its host-adapted variants. *Infect. Immun.*, **70**, 2249–2255.

Rennell,D. *et al.* (1991) Systematic mutation of bacteriophage T4 lysozyme. *J. Mol. Biol.*, **222**, 67–88.

Reuter,S. *et al.* (2014) Parallel independent evolution of pathogenicity within the genus Yersinia. *Proc. Natl. Acad. Sci. USA*, **111**, 6768–6773.

Reva,B. *et al.* (2011) Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.*, **39**, e118.

Rocha,E.P.C. *et al.* (2006) Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J. Theor. Biol.*, **239**, 226–235.

Roumagnac,P. *et al.* (2006) Evolutionary history of *Salmonella typhi*. *Science*, **314**, 1301–1304.

Shihab,H.A. *et al.* (2013) Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.*, **34**, 57–65.

Singletary,L.A. *et al.* (2016) Loss of multicellular behavior in epidemic African nontyphoidal *Salmonella enterica* Serovar Typhimurium ST313 strain D23580. *MBio*, **7**, e02265.

Sing,T. *et al.* (2005) ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**, 3940–3941.

Suzek,B.E. *et al.* (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, **31**, 926–932.

Thomson,N.R. *et al.* (2008) Comparative genome analysis of *Salmonella enteritidis* PT4 and *Salmonella gallinarum* 287/91 provides insights into evolutionary and host adaptation pathways. *Genome Res.*, **18**, 1624–1637.

Tian,W. and Skolnick,J. (2003) How well is enzyme function conserved as a function of pairwise sequence identity? *J. Mol. Biol.*, **333**, 863–882.

Viana,D. *et al.* (2015) A single natural nucleotide mutation alters bacterial pathogen host tropism. *Nat. Genet*, **47**, 361–366.

Yang,Z. and Bielawski,J.P. (2000) Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.*, **15**, 496–503.

Yang,Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, **13**, 555–556.

Yue,M. *et al.* (2015) Allelic variation contributes to bacterial host specificity. *Nat. Commun.*, **6**, 8754.

Technical details for: A profile-based method for measuring the impact of genetic variation

Nicole E. Wheeler[1]*, Lars Barquist[2], Fatemeh Ashari Ghomi[1], Robert Kingsley[3], Paul P. Gardner[1,4]

Abstract
In the following we provide some mathematical justification for the Delta bitscore metric that we evaluate in the accompanying manuscript.

Keywords
genome variation — genotype — phenotype

[1] *School of Biological Sciences, University of Canterbury, Christchurch, New Zealand.*
[2] *Institute for Molecular Infection Biology, University of Wuerzburg, Wuerzburg, Germany.*
[3] *Institute of Food Research, Norwich Research Park, Norwich, Norfolk, United Kingdom.*
[4] *Biomolecular Interaction Centre and the Bio-Protection Research Centre, University of Canterbury, Christchurch, New Zealand.*
***Corresponding author**: nicole.wheeler@pg.canterbury.ac.nz

Introduction

In this document we compare the mathematics between previously published profile HMM based methods for quantifying the likely phenotypic significance of genetic variation and our approach. Namely, the *logR.E-value* method [1], *FATHMM* [2, 3, 4] and *DBS* (this study).

1. Methods

1.1 logR.E-value

Clifford *et al.* (2004) suggest using the following measure to estimate the significance of a genetic variant:

$$\log R.E = \log_{10}\left(\frac{E-value_{var}}{E-value_{can}}\right) \quad (1)$$

Where $E-value_{var}$ and $E-value_{can}$ correspond to the expectation value derived from HMMER matches (to the same model) for a variant (*var*) and canonical (*can*) protein sequence.

$E-values$ are generally estimated by fitting an exponential distribution to an empirical (usually simulated) distribution. I.e.

$$E-value = \kappa MN e^{\lambda x} \quad (2)$$

Where x is the bit-score for a match between a profile HMM and a sequence, MN is the product of the database size and the model length and, finally κ and λ are parameters that ensure the intercept with the y-axis is correct and that the curve matches an empirical distribution .

In a breakthrough theoretical paper by Sean Eddy [5], he showed that the most computationally expensive parameter to estimate (λ) is a constant i.e. $\lambda = \ln(2)$.

Thus Equation 1 can be rewritten as:

$$\begin{aligned} \log R.E &= \log_{10}\left(e^{\lambda x_{var}}\right) - \log_{10}\left(e^{\lambda x_{can}}\right) \quad (3)\\ &= (x_{var} - x_{can}) * \log_{10}(e^{\lambda}) \\ &= DBS * constant \end{aligned}$$

If the base for the exponential and the logarithms had been equal, then *constant* the constant would equal λ. In either case, a constant multiplied by the difference between two bitscores is all that remains.

1.2 FATHMM

Shihab *et al* (2013) define the following unweighted measure for estimating the significance of a single non-synonymous SNP (the weighted version is trained to discriminate human disease from polymorphic variation, therefore is not directly comparable to our general approach). Their metric is a logit or log-odds value, comparing the emission probability of the wild-type variant (P_w) and a mutant variant (P_m) when the mutant is a single, non-synonymous point mutation (i.e. not a multiple point mutations or indels):

$$\begin{aligned} unweighted &= \ln\left(\frac{\frac{P_m}{1-P_m}}{\frac{P_w}{1-P_w}}\right) \quad (4)\\ &= \ln\left(\frac{P_m}{P_w}\right) + \ln\left(\frac{1-P_w}{1-P_m}\right) \quad (5)\\ &\approx DBS + \ln\left(\frac{1-P_w}{1-P_m}\right) \quad (6) \end{aligned}$$

The value $1-P_w$ and $1-P_m$ can be re-written as the following summation:

$$1 - P_w = \sum_{i \in amino-acids, i \neq w} P_i \quad (7)$$

$$1 - P_m = \sum_{j \in amino-acids, j \neq m} P_j \quad (8)$$

Equations 7&8 share 18 terms (for each of the 20 amino acids, less the ones corresponding to the wild-type (w) and mutant (m) variants. Therefore, $1 - P_w \approx 1 - P_m$ for most realistic biological results. As a consequence, the second term of Equation 6 is approximately zero (or at least, modest in comparison to the first term when there is a large difference between P_w and P_m). Therefore a difference between bitscores is the term that dominates Equation 4 (see the discussion below).

1.3 Delta bitscore (DBS)

Using the same nomenclature as above, we define delta bitscore (DBS) as:

$$DBS = (x_{var} - x_{can}) \quad (9)$$

The bitscore (x) for an HMM is defined as a product log of probability ratios [6]:

$$x = \log_2 \left(\frac{P(seq|M)}{P(seq|N)} \right) \quad (10)$$

Where M is a profile model derived from a sequence alignment. M generates and scores sequences based upon how likely they are to have generated by the same process as those in the sequence alignment. N is a null model, that generates and scores sequences based upon how likely they are to have generated by a random process.

Therefore, *DBS* can be re-written as:

$$DBS(seq_{var}, seq_{can}) = \log_2 \left(\frac{P(seq_{var}|M)}{P(seq_{var}|N)} \right) - \log_2 \left(\frac{P(seq_{can}|M)}{P(seq_{can}|N)} \right) \quad (11)$$

$$\approx \log_2 \left(\frac{P(seq_{var}|M)}{P(seq_{can}|M)} \right) \quad (12)$$

If we make the simplifying assumption that the null models for $P(seq_{var}|N)$ and $P(seq_{can}|N)$ are approximately equal (i.e. equal length and amino acid composition). Therefore, the first term of Equation 5 and Equation 12 are, in most situations, equivalent.

2. Discussion

As a consequence, the measures used by the log*R.E-value* (Equation 3) and the *FATHMM* (Equation 6) approach are approximations to the more direct estimation of significance, *DBS*. In the case of *FATHMM*, only single point mutations are considered, missing the wealth of variation due to insertions, deletions, multiple SNPs and and other larger-scale variants.

Consequently, *DBS* is a direct measure of the potential impact of genetic variation, that can be used on small as well as large and complex variants. We propose that this metric can be used to evaluate both population variation as well as variation between species. The mean of the distribution should be approximately zero, while the variance will increase with increasing phylogenetic distance (and different levels of selection).

One factor that may have an undue influence on *DBS* is in the rare cases where the optimal alignment between a the profile and the variant and the profile and the canonical sequence differ. For example, *HMMER3* currently only has a local mode (i.e. no "glocal" option). Therefore, splitting matches can happen, also slipped alignments also occur, particularly for repetitive sequences.

One way to mitigate these possibilities is to use *Forward Scores*, which rather than reporting just the value for an optimal alignment, reports instead the sum of all possible alignments between a query sequence and the profile model.

References

[1] R J Clifford, M N Edmonson, C Nguyen, and K H Buetow. Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. *Bioinformatics*, 20(7):1006–14, May 2004.

[2] H A Shihab, J Gough, D N Cooper, P D Stenson, G L Barker, K J Edwards, I N Day, and T R Gaunt. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden markov models. *Hum Mutat*, 34(1):57–65, Jan 2013.

[3] H A Shihab, J Gough, D N Cooper, I N Day, and T R Gaunt. Predicting the functional consequences of cancer-associated amino acid substitutions. *Bioinformatics*, 29(12):1504–10, Jun 2013.

[4] H A Shihab, J Gough, M Mort, D N Cooper, I N Day, and T R Gaunt. Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. *Hum Genomics*, 8:11, 2014.

[5] S R Eddy. A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PLoS Comput Biol*, 4(5):e1000069, May 2008.

[6] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis*. Press, Cambridge U., 1998.

Chapter Three | Draft: Genetic drift causes the accumulation of deleterious mutations in small populations of bacteria

Preface

A key part of pathogen evolution appears to be the loss of function of a subset of ancestral genes, through a combination of adaptive loss and loss due to drift. We saw a great deal of loss-of-function mutations in the pairwise comparison we focussed on in Chapter Two, and wanted to look further at the contributions of population bottlenecks and genetic drift (which are often experienced by host-adapted pathogens) to loss of gene function.

A mutation accumulation experiment presented an opportunity to examine loss of protein function due predominantly to drift, and the effects this has on fitness. In order to characterise the loss of protein function that occurs due to drift, we examined changes in the genome sequences of ten lines of *Escherichia coli* evolving in parallel that had been subjected to daily single-colony bottlenecks. We assessed the fitness of both bottlenecked lines, and control lines subjected to a less severe genetic bottleneck, compared to their ancestor.

We found that the loss of gene function observed in control lines showed greater variability across COG categories, suggesting that some functions were being selectively lost while some were being selectively maintained, while in the bottlenecked lines genes lost their function in a more uniform pattern. We saw a significant correlation between the growth rate of different lines and loss of function of genes involved in cellular energy production. Overall we found that selective loss of gene function and avoidance of highly disruptive mutations such as protein truncating mutations were characteristic of growth under stronger selection.

Contributions

Alicia Lai performed the experimental work, including whole-genome sequencing, and calculated growth and overall mutation rates. I performed the sequence analysis and most of the statistical analysis. Anthony Poole and Alicia Lai designed the original experiment and Anthony Poole supervised the project. Paul Gardner provided feedback on the manuscript.

Genetic drift drives the accumulation of deleterious mutations in small populations of bacteria

Alicia Lai[1], Nicole Wheeler[1], Paul Gardner[1], Anthony Poole[2]

1. School of Biological Sciences, University of Canterbury, Christchurch, New Zealand.
2. School of Biological Sciences, Auckland University, Auckland, New Zealand.

Abstract

Sequence evolution over time can be attributed to a combination of genetic drift and selection determining which alleles persist in a population and which are filtered out. To identify adaptive mutations in an evolving population, we must first rule out the possibility that a genetic variant has become established as the result of drift. In order to do this we must have a clear understanding of what genomic evolution under a process of drift looks like, and how this differs from selection. To address this, we have performed a mutation accumulation experiment in hypermutator *Escherichia coli* populations that have either been subjected to daily single colony bottlenecks, or subjected to serial transfer of a larger subset of the population, for 100 days. As expected, we observed a greater accumulation of mutations over time and a concurrent reduction in fitness in severely bottlenecked populations, consistent with Muller's ratchet. We also identified loss of function mutations in a range of protein coding genes over long term culture, and an increase in doubling time associated with the accumulation of deleterious mutations in genes involved in energy production. This experiment presents a contrast between genomes evolving under strong and weak genetic drift, allowing us to identify characteristic genomic changes associated with evolution under relaxed selection.

Introduction

Comparative genomics has been a boon to our understanding of bacterial pathogens (Land et al., 2015). We have identified mutations unique to pathogenic lineages that have offered new insights into the biology of host-pathogen interactions as well as promising new drug targets (Chawley et al., 2014; Langridge et al., 2015; Miesel et al., 2003). However, an ongoing criticism of comparative genomics studies has been that interpretation of results tends to be from an adaptationist perspective - that the changes we see persist over time must serve a purpose in promoting the overall fitness of the organism (Koonin, 2016). Often neglected is the accumulation of neutral or even deleterious mutations as a result of drift.

When we see change in the genome over time, our null hypothesis for the process underlying the change should be neutral evolution (Koonin, 2016; Nei, 2005). But what do changes in the genome resulting from a purely neutral process of evolution look like, and what can distinguish them from changes driven by selection?

Mutation accumulation (MA) experiments are an effective way to examine the effects of genetic drift on populations of bacteria. These experiments involve periodically applying a genetic bottleneck to a population, reducing the effective population size in order to eliminate the variation needed for natural selection, and drive evolution to occur more closely to a process of pure genetic drift (Halligan & Keightley, 2009). This will cause mutations to accumulate at the rate at which they happen, regardless of fitness effects, apart from lethal or highly deleterious mutants (Barrick & Lenski, 2013). Because relatively more deleterious mutations than beneficial mutations occur over time (Eyre-Walker & Keightley, 2007), the unidirectional ratchet-like accumulation of mutations will result in the build-up of deleterious mutations and loss of fitness, a process known as Muller's ratchet (Felsenstein, 1974; Muller, 1964). Thus, in the absence of recombination, assuming reversion mutations are rare, no individual can produce offspring with fewer deleterious mutations that it has itself (Gabriel et al., 1993). Such mutation accumulation experiments have been proposed as an effective way of distinguishing between beneficial and deleterious mutations, as under conditions of drift, both types of mutations accumulate at the rate at which they occur, while under conditions of stronger selection, deleterious mutations will be under-represented (Barrick & Lenski, 2009; Tenaillon et al., 2016).

Several studies performed in the past have noted a loss of fitness and reduction in gene function over time in lineages of *Escherichia coli* subjected to repeated single-cell bottlenecks (Funchain et al., 2000; Kibota & Lynch, 1996). This study represents an opportunity to apply next generation sequencing techniques to examine loss of fitness at the genomic level and to identify those mutations that contribute most to a reduction in fitness. The impact of mutations under both conditions has been examined in the past in *E. coli* using dN/dS to assess the functional impact of mutations (Wielgoss et al., 2013). However, a criticism of dN/dS and other methods for identifying adaptive change is that they have low sensitivity for detecting selection in recently diverged members of the same species (Mugal et al., 2014; Rocha et al., 2006), so we have performed a similar analysis here, but using a recently developed method called delta-bitscore (DBS) (Wheeler et al., 2016) in order to provide a more accurate assessment of the balance of neutral and deleterious mutations

under the two conditions. This approach incorporates information on the effects of truncations, frameshifts, indels and substitutions on protein function, to measure the difference in the accumulation of deleterious mutations across the two conditions. Using this approach we hope to identify key mutations and link them to the overall fitness of our bottlenecked lineages. It is difficult to distinguish adaptive driver mutations from non-adaptive passenger mutations using a single line (Tenaillon et al., 2016), so we used ten parallel lines in order to identify common themes.

We examined whether there was a strong correlation between loss of fitness and loss of protein function, as a way of assessing whether fitness loss resulted more from many deleterious mutations accumulating, or a small number of highly deleterious mutations affecting critical processes. We found that while the sum of deleterious mutations was a poor predictor of fitness, bottlenecked lines with lower fitness carried more deleterious mutations in genes with central roles in cell functioning. In addition, we observed a more heterogenous pattern in the accumulation of deleterious mutations across COG categories in control lines compared to bottlenecked lines, indicating greater selectivity in the cellular functions that were lost. These findings indicate that our mutation accumulation lines were able to accumulate a range of deleterious mutations growing on rich media, but that mutations involving key cellular functions were underrepresented in populations under stronger selection.

Method

Using fitness measurements and whole-genome sequencing, we examined the evolutionary dynamics of ten independent mutation accumulation lines of a hypermutator *E. coli* population growing on nutrient rich solid media. These ten populations of *E. coli* have been subjected to 100 single-colony bottlenecks. As a control, we grew five lines of hypermutator *E. coli* under serial passage conditions designed to minimise population bottlenecking while preventing population collapse due to competition for nutrients. We performed whole-genome sequencing to investigate the significance of genetic drift in driving the phenotypic and genotypic evolution of *E. coli*. To assess the impact of mutations on protein function, we used delta-bitscore to identify putative losses of function across the populations. This allowed us to predict which mutations may have contributed to the loss of cell fitness that was observed over successive generations.

Strains and media

All chemicals were purchased from Sigma-Aldrich Co. unless otherwise specified. All oligonucleotides were synthesised by Integrated DNA Technologies. *E. coli* B strain REL606 was obtained from T. Cooper (University of Houston, Texas). REL606 and REL606-derived strains were grown at 37°C in Luria Bertani (LB) media (Oxoid). For solid media, bacteriological agar (Oxoid) was added to a final concentration of 1.5% w/v. All the experiments were conducted in the presence of antibiotics at the following concentrations: Streptomycin, 100µg/mL and Ampicillin, 100µg/mL (Peptides International).

Introduction of a dominant mutator allele

A pGEM-T Easy (Promega) plasmid bearing a dominant mutD5 mutation in the dnaQ gene, which encodes the 3'-5' proofreading exonuclease of DNA polymerase III holoenzyme was introduced into REL606 by transformation. The spontaneous mutation rate of *E. coli* is approximately 1 X 10^{-3} per genome per replication, and this rate may increase 10^4 to 10^5 fold in the presence of the potent mutD5 (Cox & Lomax, 1976; Degnen & Cox, 1974).

Mutation accumulation experiment

A total of 10 genetically identical lineages were derived from a single glycerol stock of REL606 that contained a pGEM::mutD5 plasmid. Single colonies of each individual lineage were randomly picked, excised from the agar, re-suspended in 15% glycerol and streaked onto fresh Luria Bertani (LB) agar and grown at 37˚C for 24 hours (Figure 1). The pick-streak-incubate process (growth cycle) was repeated for 100 growth cycles, where the cells were maintained on LB agar plates supplemented with streptomycin and ampicillin. To ensure random selection of colonies, the last single colony of the streak, regardless of size, was selected. Five non-bottleneck control lines were also propagated through daily streaking of 100 uL of culture (agar was washed with 1 mL of 1X PBS) onto fresh LB agar. A glycerol stock of each MA (bottlenecked) lineage was prepared every day. Glycerol stocks of every tenth passage of control lineages were also prepared.

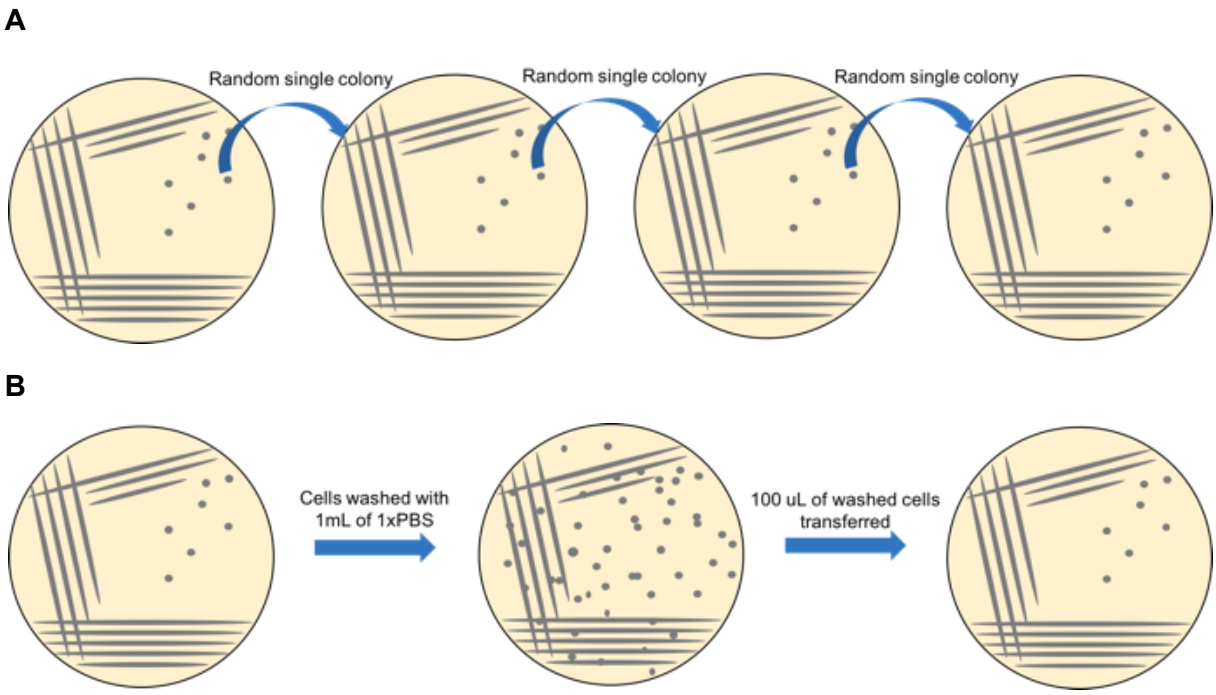


Figure 1 | Experimental design

A. The MA lines were put through single-colony bottlenecks, where a random single colony was picked and streaked onto a fresh LB agar plate. B. For the control lines, the plates were washed with 1 mL of 1xPBS and 100 uL of washed cells were streaked onto a fresh LB agar plate.

Determination of growth rate

Cultures of all of the lineages of REL606 + pGEM::mutD5 which were subjected to the MA experiment, along with the control lineages, were pre-cultured every tenth day until the cultures reached saturation in LB containing streptomycin and ampicillin. The saturated cultures were then diluted 1:100 in 2 mL of fresh LB to an OD_{595} of 0.03-0.04, and distributed into a 24-well cell culture plate. Each experimental run consisted of 3 biological replicates per sample and a negative control (fresh LB). The OD_{595} of these cultures were then monitored for a period of 24-48 hours at 37˚C (with shaking at 200 rpm), taking OD_{595} measurement every 6 minutes using a FLUOstar Omega Microplate Reader (BMG Labtech). The growth rates for each line (averaged across replicate cultures) were determined as the minimum doubling time taken over a 30-minute interval.

Whole genome sequencing

To quantify the changes that occurred at the genomic level over the course of the experiment, we sequenced the REL606 + pGEM::mutD5 ancestor line to produce a

reference genome. In addition, at the end of 100 single-colony transfers, the bottlenecked and control lineages were streaked to single colonies on LB agar. Single colonies were then used to inoculate LB liquid media. Genomic DNA was isolated using the Wizard Genomic DNA Purification Kit (Promega), and quantified using the Nanodrop 1000 Spectrophotometer and Qubit 2.0 Fluorometer. Sequencing was carried out by Macrogen Korea using an Illumina MiSeq platform with 2x250bp paired end reads. All 15 evolved samples were multiplexed in a single Illumina MiSeq Lane, resulting in 2.1 Gb of data across 9.3 million reads. The sequence data was of high quality, with 87% of bases having a Q30 value or higher. Raw sequencing data were analysed using an in-house pipeline. Briefly, the sequencing reads were processed using AdapterRemoval (Lindgreen, 2012) to remove low quality reads and adapter sequences. Following cleaning of the reads, the reads of the REL606 + pGEM::mutD5 ancestor were mapped to the REL606 genome (NC_012967.1) using Bowtie2 (Langmead & Salzberg, 2012) (using default parameters, specifying haploid genomes). The reads of the evolved lineages were then mapped to the ancestral genome. The mapping gave a sequencing depth of 32±20x (±SD). Genotyping was then carried out using SNPest (Lindgreen et al., 2014). The effects of SNPs on coding variants were predicted using Geneious v9.1.3 (Kearse et al., 2012).

Delta-bitscore analysis

The aforementioned sequenced, mapped and genotyped genomes were annotated and translated to whole proteome sequences using Geneious v9.1.3 (Kearse et al., 2012) and these protein sequences were used for delta-bitscore (DBS) calculations. DBS analyses utilise a profile Hidden Markov Model (HMM)-based approach which captures information on the expected frequency of occurrence of different amino acids, insertions, and deletions across an alignment of protein sequences (Wheeler et al., 2016). DBS is an indication of how well the evolved protein sequence fits the sequence variation modelled by an HMM, relative to an ancestral sequence. Mutations observed in highly conserved regions are likely to receive a higher score than those in non-conserved regions, and would thus give us an indication of the severity of the mutations we are seeing, and whether these are likely to be impacting protein function.

HMM profile models for gamma-proteobacterial protein sequences were retrieved from the EggNOG database (Huerta-Cepas et al., 2016). Each of our protein sequences were aligned to their respective profile HMM using HMMER3.0 (http://hmmer.org) to produce bitscore values, and by subtracting the bitscores of the ancestor protein from that of the evolved

protein, a measure of functional divergence between the proteins is produced, termed DBS (Wheeler et al., 2016). DBS was determined for all of the MA experiment lineages using this method, with a high DBS indicating a strong deviation from modelled sequence constraints for that protein, indicative of a potential loss of function. Based on benchmarking studies, a conservative scoring cutoff for identifying mutations that result in a measurable impairment of protein function *in vitro* is a DBS of 5 or above (Wheeler et al., 2016), so we used this criterion to quantify the number of genes that had incurred loss-of-function mutations.

Results

Using fitness measurements and whole-genome sequencing, we examined the evolutionary dynamics of ten independent mutation accumulation lines of a hypermutator *E. coli* population growing on nutrient rich solid media. After 100 single-colony bottlenecks over approximately 4,000 generations, we found that severely bottlenecked lineages showed lower fitness than their ancestor and weakly bottlenecked controls, and have accumulated deleterious mutations in proteins spanning a wider range of cellular functions.

An overall loss of fitness was observed

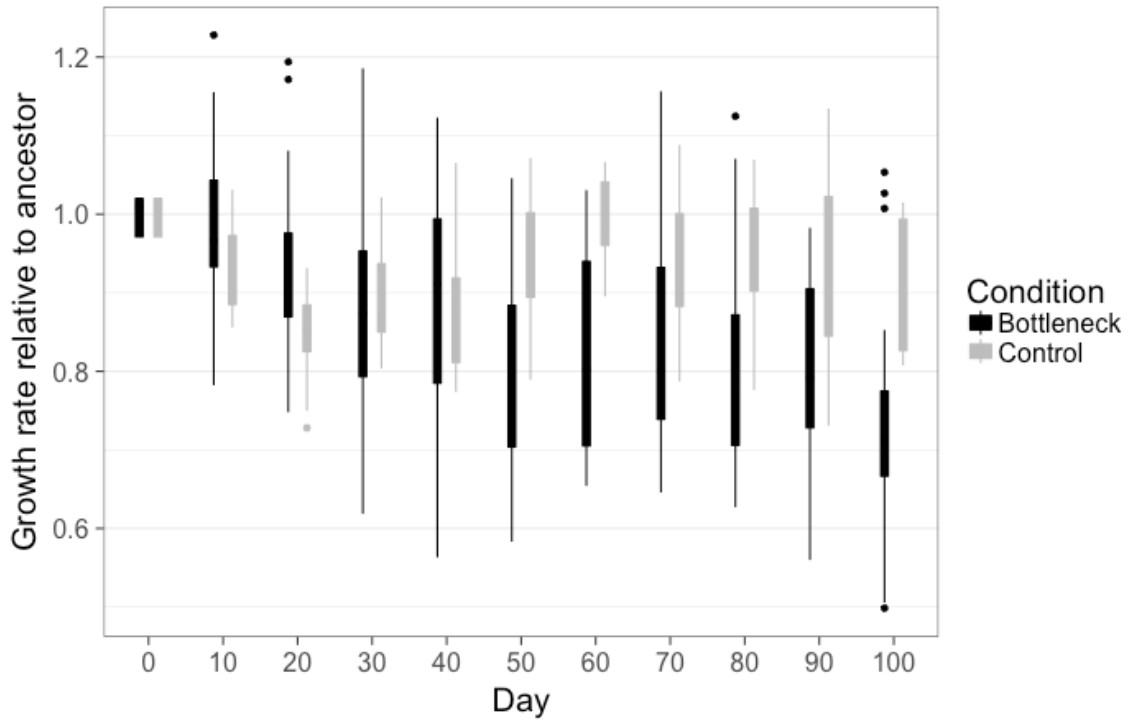


Figure 2 | The average relative growth rates of the control and bottlenecked lineages decreased with time compared to the ancestor

Growth rate was measured at ten day intervals. The average relative growth rate of the controls (n=5) and bottlenecked lines (n=9) decreased with time. Values above 1 represent

an increase in growth rate while values below 1 represent a decrease in growth rate compared to the ancestor. Growth rates for BN100.1 have been excluded because they deviate strongly from those observed in other lines and could not be measured past day 40.

During the course of our MA experiment, the bottlenecked lineages showed a decrease in fitness (measured as doubling time) relative to the ancestor (Figure 2, 35.4-53.7 min compared to 32.7 min calculated for the ancestor). The control lineages also showed a loss in fitness compared to the ancestor (Figure 2, 32.7-39.9 min compared to 32.7 min), but had higher fitness than the bottlenecked lines (P-value = 4.05×10^{-5}, Wilcoxon rank sum). An exception to this trend was BN100.10, which showed a gain in fitness compared to the ancestor (doubling time of 31.7±0.4 min).

One line developed additional hypermutator alleles and succumbed to mutational meltdown

In addition to monitoring growth rates, we also visually inspected the colony size of each lineage daily. We observed a reduction in the colony size of BN100.1 from day 10 onwards (Figure 3). Measured growth rates were consistent with the observed reduction in colony size, with doubling time increasing from 30.6±0.6 min to 82.0±3.4 min (n=3, ±SE) between days 0 and 40. This bottlenecked lineage was not revivable from glycerol stocks after 50 days. Upon whole-genome re-sequencing, we found that the introduced *mutD5* mutator allele remained in all the lineages. In addition, mutations to the *mutS*, *mutT* and *mutY* genes were observed in BN100.1. Concurrent with the rapid decline in colony size after day 10, we saw a rapid increase in the number of total mutations, and an increase in the number of mutations in genes with COG annotations relating to cell growth (Supplementary Table 1). A single mutation appeared in *mutS* at day 20 in line BN100.1, then additional mutations accumulated at days 50 and 100. We excluded the hypermutator line from subsequent tests measuring quantifiable differences in mutation rates, due to the fact that the strain had accumulated additional hypermutator mutations.

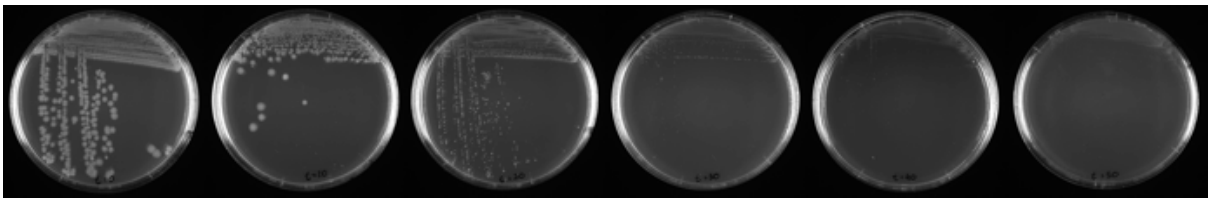


Figure 3 | Colony size decreased over time in the bottlenecked lineage BN100.1
From left to right, each plate depicts the serial single-colony bottleneck on 10-day intervals from day zero to day 50. The average colony size decreased from day 10 onward until the end of the experiment.

Whole-genome sequencing reveals greater accumulation of mutations in bottlenecked lineages

We performed whole-genome sequencing on the 100 day bottlenecked and control lines to assess the differences in mutation accumulation between the two conditions. The bottlenecked lineages, excluding BN100.1 (which accumulated approximately 13-fold more mutations compared to the other bottlenecked lineages), accumulated an average of 351±124 (±SD) mutations, while the control lineages accumulated an average of 232±48 (±SD) mutations (Supplementary Table 2). Thus, on average the bottlenecked lineages accumulated more mutations than the control lineages (*P*-value = 0.019, Wilcoxon rank sum). When we examined the number of mutations of different classes that occurred in coding sequences, we found significantly more noncoding, synonymous, nonsynonymous, and protein truncating mutations in the bottlenecked lines (Supplementary Table 3). Due to the difference in mutation rates we observed across the two lines, we corrected for both noncoding mutation rate and synonymous mutation rate where applicable, and found that protein truncating mutations were over-represented in the bottlenecked lines after correction for both (*P*-value = 0.0036).

Widespread loss of protein function occurred under both conditions

Given the observed loss of fitness in the bottlenecked lineages, we investigated whether the accumulation of deleterious mutations in protein-coding genes could have contributed to loss of fitness using a recently developed approach called delta-bitscore (DBS, (Wheeler et al., 2016)). First, we investigated whether generalised loss of protein function was correlated with an increase in doubling time. This was measured by calculating DBS values for all proteins in the evolved lineages, and taking the sum across all protein coding genes for each line. We observed high ∑DBS values in all lineages compared to the ancestor (Supplementary Table 4), indicating the accumulation of many deleterious mutations. There

was no significant difference in ∑DBS between bottlenecked and control lines (P-value = 0.254, Wilcoxon rank-sum), however the number of deleterious mutations in each strain (DBS>5) was significantly greater in the bottlenecked lines than in the control lines (mean = 19.4 vs 9.6, P-value = 0.006, Wilcoxon rank sum test).

When we plotted DBS against COG category for each of our lines, we observed that the line that succumbed to mutational meltdown had accumulated deleterious mutations across most COG categories (Supplementary Figure 1). The other bottlenecked lines also accumulated deleterious mutations across a range of COG categories, but the predicted severity was less in general (Supplementary Figure 1). This large discrepancy in high scoring mutations is predominantly due to the large number of protein truncating and frameshift mutations that occurred in BN100.1 (N = 211 and 258, respectively). Notably, the DBS distribution across COG categories in the bottlenecked line with higher fitness than the ancestor was similar to that of the control lines, but with more mutations judged to be deleterious (Supplementary Figure 1).

In order to test whether genes associated with a particular function were preferentially preserved or degraded in our experimental lines, we performed Fisher's exact tests to test for a greater accumulation of deleterious mutations in genes that belonged to each COG category compared to those that did not (Figure 4, Supplementary Table 5). The number of deleterious mutations that accumulated in each COG category over the 100 day period of the experiment was low, resulting in little power to detect significant enrichment, and as a result, no COG categories showed significant enrichment for deleterious mutations after correcting for multiple testing. Overall, the control lines showed greater variance in the odds ratios for the accumulation of deleterious mutations across COG categories (variance = 5.4 in control lines, compared to 0.3 for bottlenecked and 0.2 for mutational meltdown conditions), which can be partly attributed to lower numbers of deleterious mutations in control lines causing higher variability in odds ratios, but is also suggestive of preferential loss of function of genes that are expendable in rich media conditions.

Consistent with this hypothesis, bottlenecked lines showed a greater accumulation of deleterious mutations in energy production and nutrient transport/metabolism genes relative to the control lines, while the control lines showed a greater accumulation of deleterious mutations in cell membrane, intracellular trafficking and defence mechanism genes relative to bottlenecked lines, functions which are all likely to be redundant in our experimental

conditions. Our bottlenecked lines showed an under-representation of deleterious mutations in genes controlling translation and ribosome structure, and the nine bottlenecked lines that did not succumb to mutational meltdown showed an under-representation of deleterious mutations in genes involved in cell cycle control, indicating that even under strong genetic drift mutations encoding critical cell functions are unlikely to reach fixation in the population.

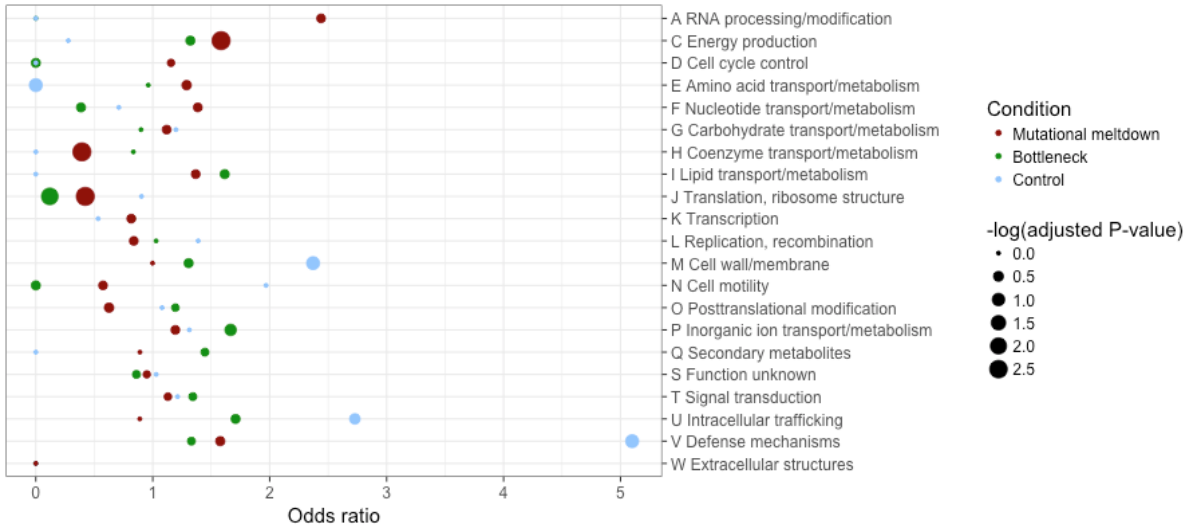


Figure 4 | Enrichment of deleterious mutations across COG categories under three experimental conditions

Odds ratios for Fisher's exact tests for enrichment of deleterious mutations in COG categories relative to other genes are shown, with the statistical significance of the result indicated by point size. The "mutational meltdown" condition corresponds to line BN100.1.

Fitness is associated with the functioning of a subset of proteins

We observed lower fitness in general in the bottlenecked lines compared to the control lines, however we also observed high variability in fitness within lines from each condition (Supplementary Table 4). To examine whether the observed difference in fitness between experimental lines could be attributed to loss of protein function, we performed a Pearson's correlation test. Neither ∑DBS nor the sum of deleterious mutations in each line correlate with the loss of fitness measured as doubling time (R^2 = 0.02, *P*-value = 0.48; R^2 = 0.23, *P*-value = 0.21). We hypothesised that while overall protein function was not significantly associated with fitness, loss of function in genes related to energy production might be, due to the fact that these genes are among the most likely to impact fitness and showed an enrichment of deleterious mutations in bottlenecked lines. As expected, we found a significant correlation between ∑DBS for genes associated with energy production (COG category C) and doubling time (R^2 = 0.61, P-value = 0.01).

Discussion

We found that under conditions of both strong and weak genetic bottlenecking, hypermutator populations of *E. coli* accumulated a range of mutations, including mutations that were predicted to be deleterious to protein function. The range of cellular functions affected by these loss of function mutations showed greater restrictions under conditions of stronger selection, indicating that while widespread loss of function is permissible on rich media, cellular functions that affect fitness are preserved when selection can act effectively.

Quantifiable differences at the genomic level between evolution under drift and selection

At the organismal level, we saw lower fitness in the bottlenecked lines compared to the ancestor and control lines. We also saw a greater accumulation of mutations, with an enrichment in protein truncating mutations above the baseline increase in mutations seen in bottlenecked lines. We did not observe an additional purging of amino acid substitutions over and above synonymous changes, suggesting that the deleterious effects of both were indistinguishable using the sample sizes and time scale we have. This can be partly explained by the fact that many proteins will not be required for survival or fitness on a rich media (Gerdes et al., 2003), therefore many nonsynonymous changes will have minimal fitness consequences under our experimental conditions. However, the difference in rates of these two classes of mutation compared to the control indicates that a subset of both synonymous and nonsynonymous mutations were having fitness consequences in the control lines. Weilgoss et al. (2013) found that lines in their long-term evolution experiment which initially developed a hypermutator phenotype subsequently accumulated other mutations which reduced mutation rates, indicating that over long-term culture, hypermutation became a maladaptive trait and was selected against. We observed a similar phenomenon here, in spite of the strong reduction in the effectiveness of selection to promote clones with lower mutations rates.

We hypothesised that we would be able to identify molecular signatures that were enriched or depleted in bottlenecked lines relative to controls, which would allow us to identify mutations that were more likely to occur under conditions of drift than under selection. Over the 100 days that passed during the experiment, significantly more mutations accumulated in our bottlenecked lines than in our control lines, however we were able to identify few statistically significant molecular signatures that would give us meaningful ability to identify individual adaptive mutations in a novel context. We suspect this is due to the low numbers

of mutations that fit individual categories, such as indels or deleterious mutations in a specific COG category. We have identified trends that we believe are worth investigating further, particularly the association between protein truncating mutations and fitness, and the enrichment and depletion of deleterious mutations across different COG categories. The greater homogeneity in functional losses associated with drift across COG categories could be an alternative metric by which we could measure the balance of adaptive and neutral change in a genome. Our findings have been echoed by studies of obligate symbiotic bacteria, which found that while metabolic genes represent the largest group of lost functions in obligate symbionts, genes from all functional categories appear to undergo pseudogenization at similar rates (Dagan et al., 2006; Gómez-Valero et al., 2004), indicating that our findings are broadly reflective of the pseudogenization events that occur in bacteria subjected to repeated population bottlenecks.

We observed stochasticity in fitness outcomes for bottlenecked lines

While we saw a general trend of gradual accumulation of mutations and loss of fitness in our bottlenecked lines, we saw two deviations from this pattern. One was a line that developed additional hypermutator mutations and succumbed to mutational meltdown during the course of the experiment, and the other was a line that showed an increase in fitness relative to the ancestor. Notably, the high fitness bottlenecked line showed a sparsity in mutations in genes involved in cellular energy production, a genomic change associated with increased doubling time. We observed the emergence of several separate potential hypermutator alleles in BN100.1, however some of these may actually have been compensatory (Rotman & Kuzminov, 2007). These findings reflect the stochasticity in fitness trajectories that can be observed in evolution under genetic drift, a phenomenon that has been observed in other studies as well (Vogwill et al., 2016).

Reduction in fitness associated with increased genetic drift is caused by a small collection of highly deleterious mutations

Sum of DBS and the total number of proteins with deleterious mutations did not significantly correlate with doubling time, indicating that a general loss of protein function does not have a strong impact on fitness under conditions of serial culture on rich media. This is to be expected, as many genes are not required for fitness or even expressed during growth on rich media (Christen et al., 2011). This highlights a key disconnect between mutations that are deleterious at the protein level and mutations that are deleterious at the organismal level. Overall, we believe these findings indicate that fitness losses were the result of strongly deleterious mutations in a few key genes, rather than the accumulation of many mildly

deleterious mutations. This hypothesis is echoed by the results of other MA experiments, such as that of Heilbron et al. (2014), who found that more than half of the decay in fitness they observed in their lines of hypermutator *P. aeruginosa* could be attributed to the fixation of a small number of highly deleterious mutations. In assessing the functional effects of mutations and how these may have contributed to the fitness of different lines, we were unable to apply DBS to mutations in noncoding regions and synonymous mutations, which can both have an important impact on organismal fitness (Agashe et al., 2016). While we acknowledge this limitation, the findings of Heilbron et al. (2014) suggest that the majority of mutations that have strong impacts on organismal fitness are indels and deleterious mutations in core genes.

Conclusion

In this study we have characterised differences in genomic evolution under strong drift compared to effective selection. As anticipated, we observed more effective purging of deleterious mutations under stronger selection, and found evidence that suggests the functional role of genes has an impact on the likelihood of deleterious mutations accumulating under conditions of selection. With the small number of mutations that had accumulated over the time scale used in the experiment, particularly in control lines, many signatures of genomic change did not show statistically significant differences across conditions over the time scale we measured. However, our finding that deleterious mutations accumulated across a broader range of gene functional categories in severely bottlenecked lines is consistent with studies on bacteria that experience population bottlenecks under natural conditions. Overall our findings indicate that protein truncating mutations are more likely to be purged under efficient selection than other types of mutations, and that the breadth of protein functions that are lost due to deleterious mutations increases with genetic drift.

References

Agashe, D., Sane, M., Phalnikar, K., Diwan, G. D., Habibullah, A., Martinez-Gomez, N. C., … Marx, C. J. (2016). Large-Effect Beneficial Synonymous Mutations Mediate Rapid and Parallel Adaptation in a Bacterium. *Molecular Biology and Evolution*, *33*(6), 1542–1553.

Barrick, J. E., & Lenski, R. E. (2009). Genome-wide mutational diversity in an evolving population of Escherichia coli. *Cold Spring Harbor Symposia on Quantitative Biology*, *74*, 119–129.

Barrick, J. E., & Lenski, R. E. (2013). Genome dynamics during experimental evolution. *Nature Reviews. Genetics*, *14*(12), 827–839.

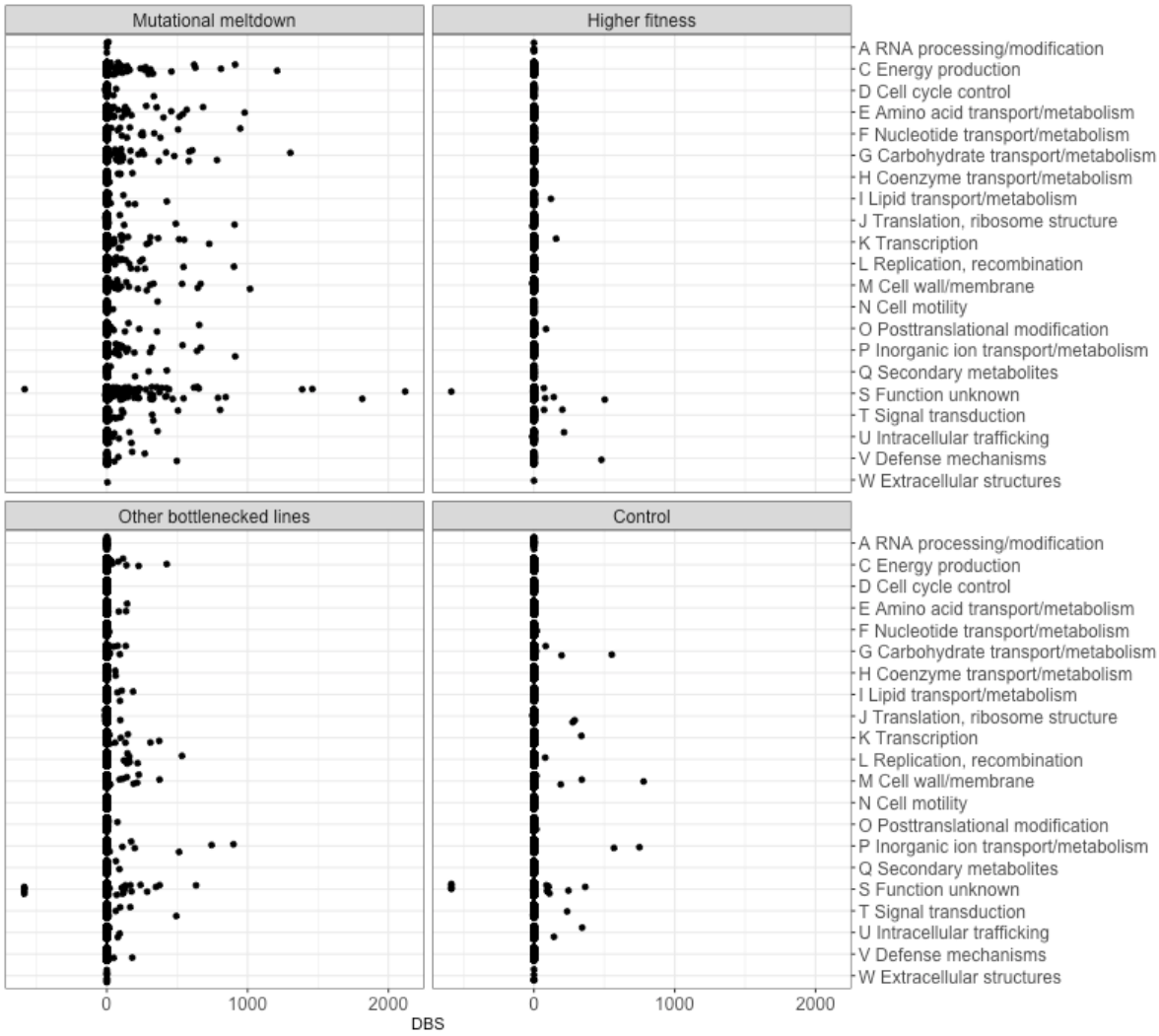
Chawley, P., Samal, H. B., Prava, J., Suar, M., & Mahapatra, R. K. (2014). Comparative genomics study for identification of drug and vaccine targets in Vibrio cholerae: MurA ligase as a case study. *Genomics*, *103*(1), 83–93.

Christen, B., Abeliuk, E., Collier, J. M., Kalogeraki, V. S., Passarelli, B., Coller, J. A., … Shapiro, L. (2011). The essential genome of a bacterium. *Molecular Systems Biology*, *7*, 528.

Cox, B., & Lomax, P. (1976). Brain amines and spontaneous epileptic seizures in the Mongolian gerbil. *Pharmacology, Biochemistry, and Behavior*, *4*(3), 263–267.
Dagan, T., Blekhman, R., & Graur, D. (2006). The “domino theory” of gene death: gradual and mass gene extinction events in three lineages of obligate symbiotic bacterial pathogens. *Molecular Biology and Evolution*, *23*(2), 310–316.
Degnen, G. E., & Cox, E. C. (1974). Conditional mutator gene in Escherichia coli: isolation, mapping, and effector studies. *Journal of Bacteriology*, *117*(2), 477–487.
Eyre-Walker, A., & Keightley, P. D. (2007). The distribution of fitness effects of new mutations. *Nature Reviews. Genetics*, *8*(8), 610–618.
Felsenstein, J. (1974). The evolutionary advantage of recombination. *Genetics*, *78*(2), 737–756.
Funchain, P., Yeung, A., Stewart, J. L., Lin, R., Slupska, M. M., & Miller, J. H. (2000). The consequences of growth of a mutator strain of Escherichia coli as measured by loss of function among multiple gene targets and loss of fitness. *Genetics*, *154*(3), 959–970.
Gabriel, W., Lynch, M., & Burger, R. (1993). Muller’s Ratchet and Mutational Meltdowns. *Evolution; International Journal of Organic Evolution*, *47*(6), 1744–1757.
Gerdes, S. Y., Scholle, M. D., Campbell, J. W., Balázsi, G., Ravasz, E., Daugherty, M. D., … Osterman, A. L. (2003). Experimental determination and system level analysis of essential genes in Escherichia coli MG1655. *Journal of Bacteriology*, *185*(19), 5673–5684.
Gómez-Valero, L., Latorre, A., & Silva, F. J. (2004). The evolutionary fate of nonfunctional DNA in the bacterial endosymbiont Buchnera aphidicola. *Molecular Biology and Evolution*, *21*(11), 2172–2181.
Halligan, D. L., & Keightley, P. D. (2009). Spontaneous Mutation Accumulation Studies in Evolutionary Genetics. *Annual Review of Ecology, Evolution, and Systematics*, *40*(1), 151–172.
Heilbron, K., Toll-Riera, M., Kojadinovic, M., & MacLean, R. C. (2014). Fitness is strongly influenced by rare mutations of large effect in a microbial mutation accumulation experiment. *Genetics*, *197*(3), 981–990.
Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., … Bork, P. (2016). eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research*, *44*(D1), D286–93.
Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., … Drummond, A. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* , *28*(12), 1647–1649.
Kibota, T. T., & Lynch, M. (1996). Estimate of the genomic mutation rate deleterious to overall fitness in E. coli. *Nature*, *381*(6584), 694–696.
Koonin, E. V. (2016). Splendor and misery of adaptation, or the importance of neutral null for understanding evolution. *BMC Biology*, *14*(1), 114.
Land, M., Hauser, L., Jun, S.-R., Nookaew, I., Leuze, M. R., Ahn, T.-H., … Ussery, D. W. (2015). Insights from 20 years of bacterial genome sequencing. *Functional & Integrative Genomics*, *15*(2), 141–161.
Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359.
Langridge, G. C., Fookes, M., Connor, T. R., Feltwell, T., Feasey, N., Parsons, B. N., … Thomson, N. R. (2015). Patterns of genome evolution that have accompanied host adaptation in Salmonella. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(3), 863–868.
Lindgreen, S. (2012). AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC Research Notes*, *5*, 337.
Lindgreen, S., Krogh, A., & Pedersen, J. S. (2014). SNPest: a probabilistic graphical model for estimating genotypes. *BMC Research Notes*, *7*, 698.
Miesel, L., Greene, J., & Black, T. A. (2003). Microbial genetics: Genetic strategies for antibacterial drug discovery. *Nature Reviews. Genetics*, *4*(6), 442–456.
Mugal, C. F., Wolf, J. B. W., & Kaj, I. (2014). Why time matters: codon evolution and the temporal dynamics of dN/dS. *Molecular Biology and Evolution*, *31*(1), 212–231.
Muller, H. J. (1964). THE RELATION OF RECOMBINATION TO MUTATIONAL ADVANCE. *Mutation Research*, *106*, 2–9.
Nei, M. (2005). Selectionism and neutralism in molecular evolution. *Molecular Biology and Evolution*, *22*(12), 2318–2342.
Rocha, E. P. C., Smith, J. M., Hurst, L. D., Holden, M. T. G., Cooper, J. E., Smith, N. H., & Feil, E. J. (2006). Comparisons of dN/dS are time dependent for closely related bacterial genomes. *Journal of Theoretical Biology*, *239*(2), 226–235.
Rotman, E., & Kuzminov, A. (2007). The mutT Defect Does Not Elevate Chromosomal Fragmentation in Escherichia coli Because of the Surprisingly Low Levels of MutM/MutY-Recognized DNA Modifications. *Journal of Bacteriology*, *189*(19), 6976–6988.
Tenaillon, O., Barrick, J. E., Ribeck, N., Deatherage, D. E., Blanchard, J. L., Dasgupta, A., … Lenski, R. E. (2016). Tempo and mode of genome evolution in a 50,000-generation experiment. *Nature*, *536*(7615), 165–170.
Vogwill, T., Phillips, R. L., Gifford, D. R., & MacLean, R. C. (2016). Divergent evolution peaks under intermediate population bottlenecks during bacterial experimental evolution. *Proceedings. Biological Sciences / The Royal Society*, *283*(1835). https://doi.org/10.1098/rspb.2016.0749
Wheeler, N. E., Barquist, L., Kingsley, R. A., & Gardner, P. P. (2016). A profile-based method for identifying

functional divergence of orthologous genes in bacterial genomes. *Bioinformatics* , *32*(23), 3566–3574.
Wielgoss, S., Barrick, J. E., Tenaillon, O., Wiser, M. J., Dittmar, W. J., Cruveiller, S., … Schneider, D. (2013). Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(1), 222–227.

Supplementary Material



Supplementary Figure 1 | DBS across different COG categories for different lineages at day 100

DBS values for each protein assigned a COG category by eggNOG are shown as points in the plot. The mutational meltdown panel corresponds to protein coding genes in line BN100.1, while the higher fitness panel corresponds to line BN100.10. The remaining 8 lines are all shown in the panel labelled “Other bottlenecked lines”. All five control lines are shown in the “Control” panel. Some genes were assigned more than one COG category, in which case they are shown as two points, one for each category. The plot shows all proteins assigned to a COG category, including those with no nonsynonymous changes, in which case DBS will be zero. The position of points is offset slightly to allow the distinction of individual mutations with similar DBS.

Supplementary Table 1 | Mutations over time for BN100.1 in all genes, and in genes from selected COG categories related to cell growth

Time point	Number of mutations					
	Total	C - energy production	D - cell cycle control	J - translation and ribosome structure	K - transcription	L - replication/ recombination
Day 0	0	0	0	0	0	0
Day 10	108	4	1	3	2	6
Day 20	599	43	4	18	17	28
Day 30	2331	146	12	49	64	80
Day 40	3129	181	19	57	82	96
Day 50	3795	199	21	67	95	104
Day 100	4803	224	17	85	109	109
Average in other bottlenecked lines at day 100 (±SD)	351 ±124	21 ±4	2 ±2	9 ±4	9 ±5	14 ±4
Average in control lines at day 100 (±SD)	231 ±54	13 ±4	1 ±1	7 ±2	9 ±3	9 ±6

Supplementary Table 2 | The number of SNPs in the bottlenecked and control lineages after 100 passages

BN represents the bottlenecked lineages while C represents the control lineages. The numbers following BN and C denote the number of passages followed by the lineage number respectively.

Lineage	Number of substitutions	Number of insertions	Number of deletions	Total number of mutations
BN100.1	4433	211	159	4803
BN100.2	371	5	9	385
BN100.3	287	10	6	303
BN100.4	216	6	7	229
BN100.5	339	7	2	348
BN100.6	578	33	35	646
BN100.7	234	6	7	247
BN100.8	323	9	7	339
BN100.9	345	18	19	382
BN100.10	268	6	4	278
C100.1	283	10	6	299
C100.2	224	7	4	235
C100.3	214	3	3	220
C100.4	229	11	14	254
C100.5	141	3	7	151

Supplementary Table 3 | Mutations accumulated in day 100 lines

		Nucleotide substitutions	Nucleotide insertions	Nucleotide deletions	Noncoding mutations	Synonymous	CDS extension	Frameshift	AA Substitution	Truncation
Line	BN100.1	4433	211	159	541	1571	8	258	2482	211
	BN100.2	371	5	9	79	91	1	9	209	5
	BN100.3	287	10	6	81	55	0	7	157	6
	BN100.4	216	6	7	102	84	0	14	248	6
	BN100.5	339	7	2	84	69	1	2	190	4
	BN100.6	578	33	35	156	149	0	35	312	6
	BN100.7	234	6	7	62	54	0	4	132	3
	BN100.8	323	9	7	78	78	0	12	166	9
	BN100.9	345	18	19	57	115	0	16	209	8
	BN100.10	268	6	4	78	67	1	8	125	8
	C100.1	283	10	6	72	57	1	12	160	0
	C100.2	224	7	4	64	59	1	8	112	1
	C100.3	214	3	3	47	59	0	4	119	0
	C100.4	229	11	14	49	51	0	8	128	2
	C100.5	141	3	7	40	29	0	2	76	2
Wilcoxon rank sum P-value (excluding BR100.1, one-sided)	Nominal	0.009491	0.2955	0.2287	0.0081	0.01635	0.6257	0.1568	0.00819	0.0015
	Corrected for noncoding mutation rate	-	-	-	-	0.6556	0.7831	0.5531	0.4469	0.0036
	Corrected for synonymous mutation rate	-	-	-	-	-	0.7831	0.5531	0.7477	0.0036
Pearson correlation P-value (one-sided)	Nominal	0.1673	0.2068	0.2318	0.4829	0.07203	0.9133	0.3247	0.0849	0.0328
	Corrected for condition (lm)								0.8369	0.8763

Supplementary Table 4 | The sum of absolute DBS and doubling time at the end evolution experiment

No growth data were collected BN100.1 as it was not revivable from glycerol stock

Lineage	∑DBS	Doubling time (minutes) (n=3, ±SE)
BN100.1	77203.1	*
BN100.2	631.7	44.1±0.3
BN100.3	1240.5	47.4±1.7
BN100.4	258.1	44.3±2.2
BN100.5	590.1	46.9±0.6
BN100.6	5779.7	39.8±1.1
BN100.7	1342.8	48.8±0.6
BN100.8	2343.8	50.2±3.6
BN100.9	1786.5	64.4±0.7
BN100.10	1699.0	31.7±0.4
C100.1	2295.1	36.8±0.9
C100.2	1724.4	32.7±0.4
C100.3	289.7	40.0±0.3
C100.4	672.1	39.3±0.3
C100.5	15.4	32.7±0.2

Supplementary Table 5 | Enrichment of deleterious mutations in bottlenecked and control lines across COG categories

Enrichment was tested using a Fisher's exact test. Counts exclude line BN100.1, as it developed additional hypermutator alleles and was no longer considered comparable. Deleterious mutations were defined as having a DBS > 5. P-values were adjusted for multiple testing using the Benjamini-Hochberg procedure.

COG	Condition	Deleterious mutation in COG	Neutral mutation in COG	Deleterious mutation in other COGs	Neutral mutation in other COGs	P-value	Odds ratio	Adjusted P-value
A RNA processing/ modification	Bottleneck	0	36	178	35816	1	0	1
	Control	0	20	49	19942	1	0	1
	Mutational meltdown	1	3	476	3484	0.401	2.439	0.731
C Energy production	Bottleneck	16	2493	162	33359	0.299	1.322	0.697
	Control	1	1393	48	18569	0.259	0.278	1
	Mutational meltdown	48	230	429	3257	0.007	1.584	0.070
D Cell cycle control	Bottleneck	0	396	178	35456	0.273	0	0.697
	Control	0	220	49	19742	1	0	1
	Mutational meltdown	6	38	471	3449	0.645	1.156	0.847
E Amino acid transport/ metabolism	Bottleneck	12	2507	166	33345	1	0.962	1
	Control	0	1399	49	18563	0.048	0	0.336
	Mutational meltdown	41	237	436	3250	0.152	1.289	0.662
F Nucleotide transport/ metabolism	Bottleneck	2	1024	176	34828	0.252	0.387	0.697
	Control	1	569	48	19393	1	0.710	1
	Mutational meltdown	18	96	459	3391	0.240	1.385	0.721
G Carbohydrate transport/ metabolism	Bottleneck	14	3107	164	32745	0.790	0.900	1
	Control	5	1729	44	18233	0.612	1.198	1
	Mutational meltdown	45	297	432	3190	0.487	1.119	0.731
H Coenzyme transport/ metabolism	Bottleneck	5	1201	173	34651	1	0.834	1
	Control	0	670	49	19292	0.414	0	1
	Mutational meltdown	7	127	470	3360	0.010	0.394	0.070
I Lipid transport/ metabolism	Bottleneck	6	758	172	35094	0.285	1.615	0.697
	Control	0	425	49	19537	0.628	0	1
	Mutational meltdown	13	70	464	3417	0.305	1.368	0.731
J Translation, ribosome structure	Bottleneck	1	1619	177	34233	0.005	0.119	0.114
	Control	2	898	47	19064	1	0.903	1
	Mutational meltdown	10	168	467	3319	0.005	0.423	0.070
K Transcription	Bottleneck	11	2652	167	33200	0.666	0.825	0.932
	Control	2	1477	47	18485	0.582	0.533	1
	Mutational meltdown	30	265	447	3222	0.352	0.816	0.731
L Replication, recombination	Bottleneck	11	2158	167	33694	0.874	1.028	1
	Control	4	1201	45	18761	0.538	1.389	1
	Mutational meltdown	25	216	452	3271	0.475	0.838	0.731
M Cell wall/membrane	Bottleneck	15	2359	163	33493	0.290	1.307	0.697
	Control	7	1311	42	18651	0.040	2.371	0.336
	Mutational meltdown	31	227	446	3260	1	0.998	1
N Cell motility	Bottleneck	0	378	178	35474	0.271	0	0.697
	Control	1	209	48	19753	0.404	1.969	1
	Mutational meltdown	3	38	474	3449	0.472	0.575	0.731030556
O Posttranslational modification	Bottleneck	8	1360	170	34492	0.555	1.193	0.832
	Control	2	757	47	19205	0.709	1.080	1
	Mutational meltdown	12	138	465	3349	0.158	0.626	0.662239775
P Inorganic ion transport/ metabolism	Bottleneck	18	2268	160	33584	0.045	1.666	0.469
	Control	4	1266	45	18696	0.552	1.313	1

	Mutational meltdown	35	217	442	3270	0.367	1.193	0.731030556
Q Secondary metabolites	Bottleneck	3	420	175	35432	0.469	1.446	0.804
	Control	0	235	49	19727	1	0	1
	Mutational meltdown	5	41	472	3446	1	0.890	1
S Function unknown	Bottleneck	38	8598	140	27254	0.481	0.860	0.804
	Control	12	4783	37	15179	0.869	1.029	1
	Mutational meltdown	110	837	367	2650	0.689	0.949	0.851
T Signal transduction	Bottleneck	8	1214	170	34638	0.401	1.343	0.804
	Control	2	677	47	19285	0.683	1.212	1
	Mutational meltdown	18	117	459	3370	0.592	1.129	0.828
U Intracellular trafficking	Bottleneck	7	839	171	35013	0.204	1.708	0.697
	Control	3	466	46	19496	0.107	2.728	0.563
	Mutational meltdown	10	82	467	3405	0.871	0.889	1
V Defense mechanisms	Bottleneck	3	456	175	35396	0.497	1.331	0.804
	Control	3	252	46	19710	0.025	5.099	0.336
	Mutational meltdown	9	42	468	3445	0.198	1.577	0.692
W Extracellular structures	Bottleneck	0	9	178	35843	1	0	1
	Control	0	5	49	19957	1	0	1
	Mutational meltdown	0	1	477	3486	1	0	1

Chapter Four | Draft: Identification of genomic changes associated with invasiveness in *Campylobacter jejuni*

Preface

An effective way to identify mutations associated with adaptation to a new niche is to observe the same changes occurring in parallel in bacteria adapting independently to the same conditions. An exciting opportunity presented itself when a collection of invasive *Campylobacter* isolates from New Zealand became available for study. Each isolate came from a different clinical case, and was suspected to be a unique instance of a *Campylobacter jejuni* isolate having acquired the ability to cause invasive infection. This was supported by comparison of the genomes of these clinical isolates with reference *C. jejuni* isolates. They were each more closely related to isolates from the cases of gastroenteritis in the United Kingdom than to each other, indicating that they had not originated from a single outbreak of invasive *Campylobacter*, but rather had diverse evolutionary origins.

Traditional comparative analyses had failed to identify any genomic features that distinguished them from their close relatives. This was an opportunity to apply delta-bitscore and test whether it had any ability to detect genomic changes associated with invasiveness that had been missed by other methods. The probability of finding the same mutation that had occurred in parallel in all ten invasive isolates was low, but there was potential to find different mutations occurring in the same gene that each contributed to adaptation to an invasive lifestyle via a similar mechanism. Traditional statistical methods failed to find any genes that showed consistently lower (or higher) functional potential in invasive isolates, but this was to be expected due to the large number of genes that could possibly be involved in adaptation to invasiveness.

Because the data were likely to contain many genes not involved in niche adaptation (resulting in ‘sparse’ data), a random forest algorithm was applied to the data to identify the best predictors of invasiveness. The predictor that was built had little value in identifying novel cases of invasive *Campylobacter jejuni*, but identified several proteins that were truncated in a number of invasive isolates, and cell shape determining proteins showing more subtle mutational patterns associated with invasiveness. These signs of adaptation

may indicate genomic changes that facilitated invasive infection, or changes following movement into an invasive niche that improved fitness.

Contributions

Other authors performed the initial analysis to find there were no identifiable differences in gene content between the isolates, I performed DBS analysis and wrote the manuscript. Paul Gardner provided feedback and guidance.

Identification of genomic changes associated with invasiveness in *Campylobacter jejuni*

NE Wheeler[1,2], PJ Biggs[3,4,5], T Blackmore[6], AD Reynolds[7], AC Midwinter[4], J Marshall[4], PP Gardner[1,2] and NP French[3,4]

1. School of Biological Sciences, University of Canterbury, Christchurch, New Zealand.
2. Biomolecular Interaction Centre, University of Canterbury, Christchurch, New Zealand.
3. Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand.
4. mEpiLab, Institute of Veterinary, Animal and Biomedical Sciences, Massey University, Palmerston North, New Zealand.
5. New Zealand Genomics Ltd (NZGL – as Massey Genome Service) Massey University, Palmerston North, New Zealand.
6. Capital and Coast District Health Board, Wellington, New Zealand.
7. AgResearch, Hopkirk Research Institute, Palmerston North, New Zealand.

Abstract

Campylobacter jejuni is the most common cause of bacterial diarrhoeal disease in the world. Clinical outcomes of infection can range from asymptomatic infection to life threatening invasive disease. This variability of outcomes for infected patients has raised questions as to whether genetic differences between *C. jejuni* isolates could lead to differences in the invasive potential of isolates, making some strains more clinically dangerous than others. In this study we compare the genomes of ten invasive *C. jejuni* isolates from New Zealand with reference isolates from the Oxfordshire surveillance project in order to assess whether there are recurring patterns in the accumulation of mutations in protein coding genes shared by these isolates that are specific to invasive strains. We identified a collection of genes that display sequence divergence patterns associated with invasive infection, including some that have been previously linked to virulence and invasiveness in *C. jejuni*. This study presents a screen for functionally important sequence variation associated with a phenotype of interest that can be applied more broadly to other datasets, to improve our understanding of the genomic changes that occur when bacteria transition to a new niche.

Introduction

Campylobacter jejuni is the most common bacterial cause of diarrhoea in the world (Kirk et al., 2015). Variability of clinical presentation of *Campylobacter* infection has been a subject

of interest for some time now (Carvalho et al., 2001), as clinical presentation can range from asymptomatic infection, to diarrhoea (Calva, 1988), to invasive infections such as meningitis and bacteremia (Calva, 1988; Goossens et al., 1986; Skirrow et al., 1993). An 11 year surveillance study conducted in England and Wales from 1981-1991 found that 1.5 out of every 1000 *Campylobacter* infections resulted in *Campylobacter* bacteremia (Skirrow et al., 1993). Of these cases, 80% were caused by *C. jejuni*. *Campylobacter* bacteraemia appears to originate from acute colitis, suggesting a progression from diarrhoeal disease to a more severe presentation in a subset of infected individuals (Skirrow et al., 1993). This wide range of disease presentations can not be explained purely by host factors, indicating that differences in the bacterial pathogen must also contribute (Carvalho et al., 2001). Over decades of study, toxin production and the ability to adhere to epithelial cells have been implicated as *Campylobacter* genetic factors that are associated with disease presentation (Fauchere et al., 1986; Ruiz-Palacios et al., 1983), as well as the presence of specific genetic loci (Burucoa et al., 1995; Carvalho et al., 2001; Konkel et al., 1997; Pei & Blaser, 1993).

Campylobacter is a major source of bacterial infections in New Zealand. The pathogen has gained more attention in the public eye after contamination of drinking water with *Campylobacter* in Havelock North in 2016 caused an estimated 5000 people (approximately one third of the town) to contract gastroenteritis (Gastro bug hit 5000, 2016). More generally, New Zealand ranks near the top in the world for its rate of *Campylobacter* infections (Lake et al., 2007), making this an infectious agent of strong relevance to public health, and suggesting that there may be lifestyle and/or genetic factors of both host and pathogen that put New Zealanders at greater risk. Given the high incidence of infections in New Zealand, finding a way to identify isolates that pose a greater threat to health through their increased ability to cause invasive infection would be a valuable improvement to our response to controlling the pathogen.

In order to assess whether there may be genetic factors associated with invasive infection in New Zealand isolates of *Campylobacter jejuni*, ten isolates of *C. jejuni* were isolated at Wellington Hospital from the blood (n=9) or joint aspirate (n=1) from patients showing signs of *Campylobacter* bacteraemia. These were compared to closely related reference strains from the Oxfordshire surveillance project (OXC), a genomic surveillance program investigating seasonal and temporal trends in *Campylobacter* that cause gastroenteritis in Oxfordshireshire, United Kingdom (Cody et al., 2012). No significant differences in overall

gene content or distribution of genes across functional categories could be identified, so we investigated whether a difference in the accumulation of mutations in orthologous genes could explain the differences in disease presentation. These strains are more closely related to reference isolates from the OXC that cause gastroenteritis than to each other, suggesting that if any genomic changes have occurred to facilitate invasiveness, they have occurred very recently. If any changes have occurred after invasive infection, little time has passed while the isolate has been living in the host for genomic changes associated with changes in selective pressures to have occurred in parallel across strains. Because of this, we believed that any signs of adaptation to an invasive lifestyle would be present in only a subset of lineages, and wished to assess whether investigation of adaptation at this time scale using such a small sample size was viable.

We performed a comparative analysis to determine whether there were any detectable, recurrent genomic changes observable in invasive isolates but not in gastrointestinal isolates that may explain differences in disease presentation. We hypothesised that we may be able to identify strong selective pressures on these invasive isolates that cause functionally significant changes to accumulate in a considerable proportion of isolates soon after the shift in lifestyle. We employed a recently developed profile hidden Markov model based method for identifying functionally significant changes in protein coding genes (Wheeler et al., 2016), then used both traditional statistical approaches, and a machine learning based approach to identify informative genes. We found that there was little crossover in the informative genes identified by both methods. Traditional statistical approaches identified a large number of false positive associations that were due to differences in annotation methods, while the machine learning based approach identified strong candidates and captured functionally related genes that were more informative of phenotype in combination than when they were considered separately.

Method

In this investigation we compared ten invasive clinical isolates of *Campylobacter jejuni* to closely related gastrointestinal strains from the Oxfordshire surveillance project in order to identify sequence differences in the genes shared by these isolates that were characteristic of invasive isolates. We employed an adaptation of a recently developed profile hidden Markov model-based approach (Wheeler et al., 2016) to quantify the functional significance of nonsynonymous changes in protein coding genes present in at least half of the isolates from each class. We used classical statistical techniques and two different supervised

classification methods to identify informative genes, then examined the value they offered in differentiating the two classes.

Sample collection

Samples were collected from Wellington hospital from 2010 to 2012. The age of patients the samples were collected from ranged from 19-89 years. 6 presented with diarrhoea, others presented with headache, prosthetic hip infection and exacerbation of chronic pulmonary disease. 9 samples were taken from blood and one from joint aspirate. 6/10 patients were subsequently treated with ciprofloxacin, one patient died, and one had had a previous invasive infection 4 years earlier. Only one person in the study showed sepsis syndrome. For each sample, sequence data for two random isolates from the same clonal complex were retrieved from the Oxfordshire surveillance project collection (Cody et al., 2012) for comparison.

DNA extraction and sequencing

Isolates were grown on Columbia horse blood agar plates (Fort Richard Laboratories, Auckland, New Zealand) in a microaerobic atmosphere (85% N_2, 10% CO_2, 5% O_2) at 42 degrees C for 24 hours in a VA500 variable atmosphere incubator (Don Whitely Scientific, Yorkshire, United Kingdom). DNA was extracted using a QIAamp DNA Mini kit (Qiagen, Hilden, Germany). The resultant DNA was then sent to the Massey Genome Service, Massey University, Palmerston North (part of New Zealand Genomics Ltd) for quality control checking, library preparation and sequencing. The isolates were sequenced on an Illumina MiSeq (250bp paired end run) to ~140x – 200x estimated coverage based on the size of the *Campylobacter jejuni* genome.

Read curation and assembly

Paired sequence reads were analysed with an in-house quality control tool that performed a number of functions: read quality analysis and visualisation (SolexaQA++ and FastQC) (Andrews, 2010; Cox, Peterson, & Biggs, 2010), PhiX removal with Bowtie2 (Langmead & Salzberg, 2012) and the SamToFastq.jar program from the Picard suite (http://picard.sourceforge.net), and adapter removal through the “fastq-mcf” program from the ea-utils suite of tools (Aronesty, 2011). In addition, the reads were analysed with FastQScreen (http://www.bioinformatics.babrham.ac.uk/projects/fastq_screen) as a further check for any potential Illumina adapters and cloning vector contamination. The filtered reads were assembled using Velvet (version 1.2.10) (Zerbino & Birney, 2008) in a paired end

mode, and all other parameters were set as default. Assemblies were performed on the reads at a range of subsets (294k, 336k, 378k and 420k) and for kmers from 245 to 55 in decrements of 10, resulting in 80 assemblies per isolate. The metrics of the assemblies were evaluated by generating rankings based on the values of four parameters: the longest assembly, the longest contig, the highest N_{50} value, and the fewest contigs. The assembly with the lowest summated ranking was used for further analysis.

Core genome generation

The contigs were annotated using the PROKKA (Seemann, 2014) annotation suite with default parameters, with all outputs being stored inside a MySQL (https://www.mysql.com) database. For a set of contigs under analysis, the amino acid predictions were extracted from the database, and gene clustering was performed using OrthoMCL (Li, 2003). Again, outputs were stored in a MySQL database. A set of orthologous genes that were present in all contigs under analysis was chosen for further investigation. The lengths of the gene in an orthologous cluster were analysed and if they were with a length range of 20% of the longest gene prediction in the cluster, then that cluster formed part of the core genome for those contigs. The sequences of the predicted genes in the cluster were then checked to see if they were different. Genes that fell into this category (same length within a cluster but different sequences, or length variation within a cluster) were then concatenated per isolate in order to form a multi-gene alignment that was then subsequently visualised in SplitsTree (Huson & Bryant, 2006).

Identification of functional divergence in orthologous proteins

For the purposes of identifying functional divergence in orthologous proteins, orthologous groups were filtered for any that contained more than one gene from the same strain (N = 80). The remaining genes (N = 2163) were scored against HMM profile models for epsilon-proteobacterial protein sequences retrieved from the EggNOG database (Huerta-Cepas et al., 2016). Comparisons were performed using the HMMER3.0 package (http://hmmer.org). The top scoring model for each gene was used, then orthologous groups were checked for consistent model matches. Groups which contained members that had different top scoring model hits were excluded from analysis (N=40), such as a group that had 15 hits to ENOG410I5B9, a bacteriophage Mu Gam like protein, and 15 hits to ENOG410I2J8, the flagellar biosynthesis protein FlhF. We also excluded orthologous groups that had no match to an eggNOG family (N=249). This reduced the number of orthologous genes tested to 1874. Furthermore, groups with representative proteins from fewer than five

invasive or ten gastrointestinal strains were also excluded, leaving 1432 orthologous gene families remaining for comparison.

Identification of informative genes using traditional univariate statistical methods

In order to test whether there were genes that showed differences in score distributions between invasive and gastrointestinal isolates that could be detected using traditional statistical methods, we identified the genes with the most extreme 5% of differences in median bitscore between the two groups, and the genes with the most extreme 5% of differences in distribution, as measured by a Kolmogorov Smirnov test, and identified the overlap in these two gene sets as potentially informative.

Identification of invasive strains using recursive partitioning

To get an indication of how easily invasive strains could be separated out from gastrointestinal strains, a recursive partitioning approach was employed to select the best genes from the training dataset to separate out the two classes. The R package “rpart” (Therneau et al., 2015) was used to build a recursive partitioning tree using the training set. Then, to compare the ease of separation to a null, the classes were randomly reassigned at the same frequencies 1000 times and used to build new trees using the training data. Ease of partitioning was tested by calculating the number of nodes (genes) required to achieve perfect separation of both classes.

Using a random forest classifier to identify key genes

Random forests are a data-mining technique that are gaining interest in the field of genome-wide association studies for their ability to deal with sparse, high-dimensional data (Pappu & Pardalos, 2014). Random forests are an ensemble of recursive partitioning trees (Breiman et al., 1984), which are a popular supervised learning strategy due to the fact that they are relatively transparent and interpretable. Although random forests are primarily designed for classification of new samples, they produce variable importance measures which indicate how informative each variable has been in separating cases according to category. These variable importance measures can be used to assess which genes may be important in the adaptation of invasive *Campylobacter* to their niche.

A random forest algorithm was used to build classification trees from sub-samples of the orthologous groups. Because the dataset was likely to be extremely sparse, meaning that most of the genes in the training dataset offer little to no predictive value in determining the

phenotype of interest, we performed feature selection first, selecting only genes that showed differences in median score between the two groups that is in the greatest 80% (|DBS| ≥ 0.4), and overlap in distributions, as measured by a Kolmogorov Smirnov test that fell in the lowest 80% of all KS statistics computed for orthologous groups (KS ≥ 0.2). This reduced the set of genes from 1874 to 94.

A random forest was built using 1000 trees, sampling 30 genes for each node. Decision trees within the forest were built using the R package “randomForest” (Liaw & Wiener, 2002), which functions as such: for each sub-sample of the strains, the gene with a score distribution that best distinguished between the two groups was selected as the first node in the decision tree, and assigned an optimal scoring cutoff for partitioning samples into daughter nodes. For each of the two groups resulting from the split, if another gene could further separate the two classes, this was included in the tree as well. For each node in the tree, a new sample of 30 genes was taken. Trees are built until no more information from the sampling of genes can improve discrimination between the two classes. Key indicator genes were then selected based on their feature importance scores, which are calculated based on the improvement in homogeneity of the two nodes that result from the split (Breiman, 2001). Again, to assess the significance of the variable importance measures, the random forest model was re-run 1000 times with the classes permuted each time, in an approach similar to that taken by Huynh-Thu et al. (2008).

Results

In this investigation we examined whether sequence differences in protein coding genes shared between invasive and noninvasive isolates of *C. jejuni* could be used to predict phenotype using traditional univariate statistical methods and two supervised classification methods, recursive partitioning and random forests. With the first supervised classification method, we assessed how easily the two phenotypes could be identified using the information on sequence variation we had available, and with the second we assessed which genes offered the most value in separating the two classes of infection outcomes. We found that the patterns of variation we identified in our study are unlikely to have strong predictive value in identifying new cases of invasive *Campylobacter jejuni*, however they are indicative of some of the selection pressures encountered during invasive infection.

Traditional comparative genomic analysis offers no explanation for differences in infection outcome

The overall genome size of the invasive isolates was similar to that of the gastrointestinal isolates examined (1.659 Mb ± 14 kb vs 1.654 Mb ± 22 kb), as was the number of predicted genes (1714+18 vs 1705±30). GC content was also indistinguishable (30.38±0.05% vs 30.38±0.06%). We found no clear indication that there were genes found only in invasive isolates and not in gastrointestinal isolates. A NeighbourNet diagram of relatedness of the strains used in our study indicated that each invasive strain was closely related to the selected reference strains, but that invasive strains from the same clonal complex were also closely related to each other (Figure 1).

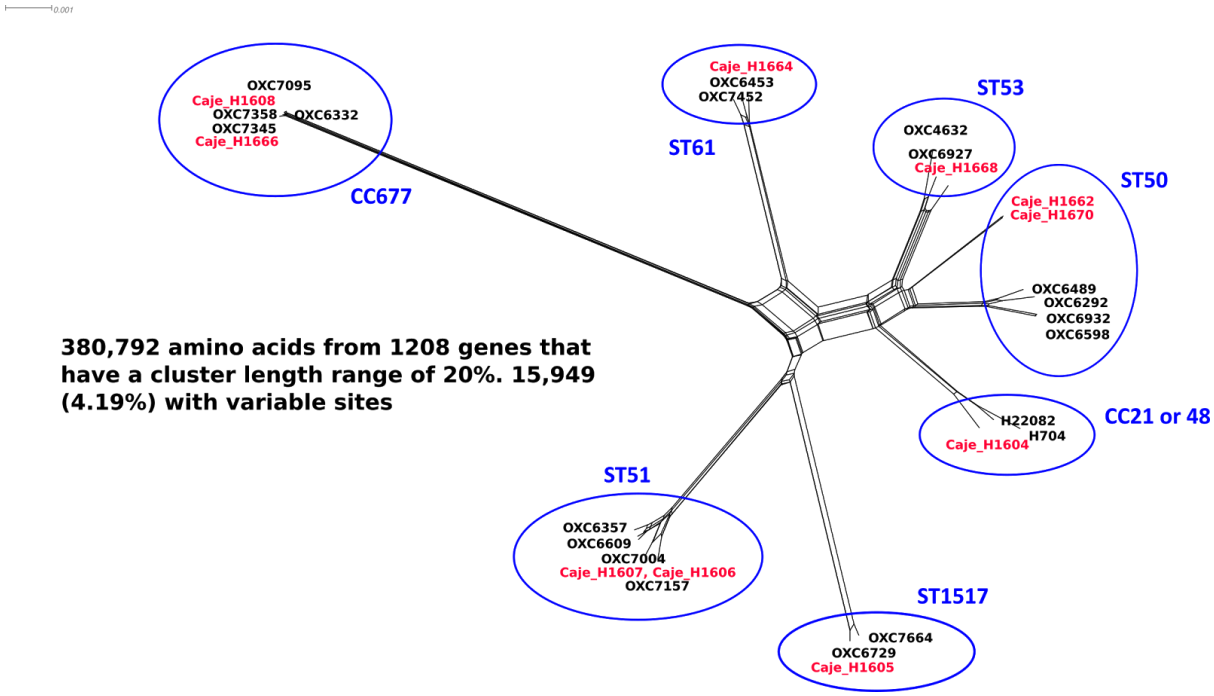


Figure 1 | NeighbourNet diagram of relatedness of strains included in the study
Invasive strains begin with the prefix Caje. Each invasive strain was paired with two gastrointestinal strains from the Oxfordshire surveillance project from the same clonal complex. Colours correspond to different clonal complexes.

Univariate statistical approaches fail to identify markers of invasive infection

In order to investigate whether any genes showed consistent signs of functional adaptation across all invasive isolates sequenced that were detectable using traditional statistical techniques, we identified the genes that showed differences in bitscore (DBS) between the

two groups that were in the most extreme 5% of values (|DBS| ≥ 0.82), as well as KS statistics in the most extreme 5% of all KS statistics computed for orthologous groups (KS ≥ 0.3). With each individual test, we would expect approximately 72 genes to qualify (taken as a percentage of total genes tested), but when we took the intersection of genes meeting both criteria, 11 genes met the two requirements. There were no genes that showed consistently higher functional potential in one group compared to the other, in fact the maximum KS value achieved for an orthologous group was 0.56. This indicates that none of these genes in isolation would be predictive of an invasive phenotype in *C. jejuni*.

Additionally, most genes identified using this approach showed differences in length between the two groups. Some of these differences appeared to be mutations resulting in frameshifts and truncations, others looked like they may have been a result of inconsistent identification of the start codon between New Zealand and Oxfordshire isolates due to differences in annotation approaches (Supplementary Table 1). We hypothesised that information from multiple genes in combination could be used to build a better predictor of infection outcome than the functional scores of any single gene. At the short timescales since the emergence of these invasive isolates, it is more likely that any sign of adaptation to an invasive lifestyle in a single gene would only have occurred in a subset of the strains, making a combinatorial approach to finding indicators of invasiveness more promising than searching for consistent trends in a gene that appear across all invasive isolates.

Distinction of invasive and gastrointestinal C. jejuni is achievable using a small set of informative genes

Two supervised classification approaches were used to classify *Campylobacter* isolates based on invasiveness and then to assess whether there is real predictive value in the genes identified as informative. The first is recursive partitioning using the “rpart” package in R (Therneau et al., 2015). In recursive partitioning, a decision tree is built to classify the strains based on invasiveness. The first gene which best splits the strains into two groups is chosen for the first node, then the data are separated according to this gene and the next best gene is selected for the resulting two groups separately. This procedure is repeated recursively until a desired level of separation is achieved. By default, this process will stop before it reaches the point of differentiating individual strains in order to avoid overfitting, but for the purpose of our investigation, the classifier was allowed to continue until the strains had been perfectly separated. We performed partitioning to perfect separation in order to get an indication of the complexity of the decision tree that would be required to perform such a

task, and compared this to the complexity of decision trees required to perfectly separate randomised classes.

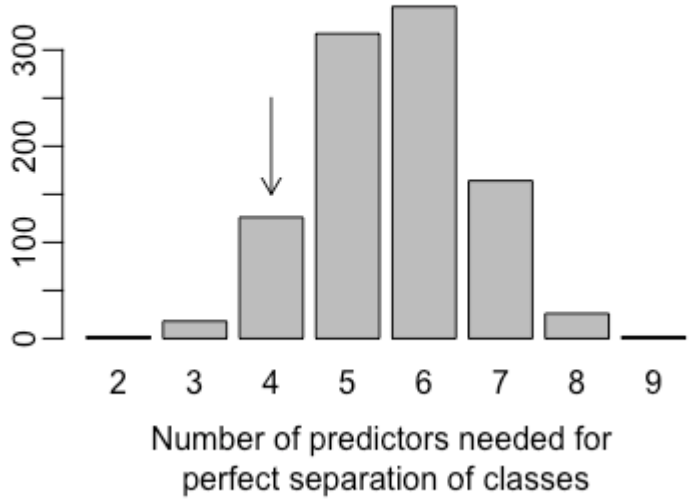


Figure 2 | Number of predictors (genes) needed to distinguish between randomised classifications across 1000 permutations

The number of predictors needed to distinguish between the real classes is indicated by an arrow.

Using the recursive partitioning approach to determine the effectiveness of classifying invasive *Campylobacter* compared to classifying random groupings showed that with real classes, four splits needed to be made (meaning four genes are required to separate all samples into their correct classes), fewer than the number required for 87% of randomly assigned classes (based on 1000 permutations) (Figure 2). Of the top identified features, few were highly correlated, suggesting they would have good ability to identify different combinations of invasive strains. The first node in the recursive partitioning tree separated out 4/10 of the invasive strains into a homogenous daughter node. Each successive node identified another two invasive samples until all samples were split into pure groups. For each node, there were several genes with similar or equal discriminatory power that could have been used in place of the gene chosen, indicating that there were multiple solutions to perfectly separating the classes. The top gene chosen for the first node was OXC6292_01615, which was also identified as one of the 11 genes that was most informative using traditional univariate statistical tests. This is to be expected for the first node in the tree, as it is selected on the basis of its ability to separate the largest number of samples correctly. In contrast, only one of the other top predictors selected for other nodes in

the recursive partitioning tree (either as a final selection or as a surrogate variable), OXC6292_01248, was identified as being independently informative in our univariate tests.

A random forest approach identifies the best combination of genes for distinguishing invasive strains

The second method we employed was a random forest algorithm. This process essentially replicated a recursive partitioning technique, but many times over. Each tree in the forest is built using a bootstrap sampling of the *Campylobacter* isolates (N = 30), and for each node in the recursive partitioning tree, a random subsampling of genes was taken, then the gene with the best ability to separate invasive and gastrointestinal isolates based on DBS is selected for that node. The aim of this process is to identify as many unique informative combinations of genes for classification purposes, then rank the individual genes based on how much they contributed to the ability of the random forest to separate invasive and gastrointestinal isolates.

Out-of-bag (OOB) estimates of error can be used to assess the accuracy of a random forest predictor. During the training of a random forest classifier, the isolates left out of training each tree during bootstrap sampling can be used to test the classification accuracy of the tree on data it has not been trained on (Breiman, 1996). Because each tree is tested on samples it has not encountered before, the OOB error estimate gives an indication of the ability of the model to detect new invasive isolates not included in the training data. Using a Random Forest approach, the OOB estimate of error was 23.3% for the real classes. Only 2% of the models built using permuted classes had error rates this low (Figure 3). However, this 23.3% overall error rate incorporated a 60% error rate in identifying invasive strains (Table 1), indicating that the model was only able to correctly identify 4/10 invasive strains when extrapolating data from bootstrapped training samples to unseen samples. This appears to be largely due to the fact that most clonal complexes in the study had one invasive representative and two gastrointestinal, so any tree that was trained without a given invasive isolate was likely to capture features unique to that clonal complex and associate them with gastrointestinal pathogens. Of the correctly identified invasive isolates (Caje_H1606, Caje_H1607, Caje_H1662 and Caje_H1670), each belonged to a clonal complex containing another invasive isolate, meaning that the trees these isolates were tested with had been trained using a closely related invasive isolate and were able to identify lineage-specific determinants of phenotype in the training data. These results indicate that

the patterns associated with invasiveness do not generalise well across the different isolates.

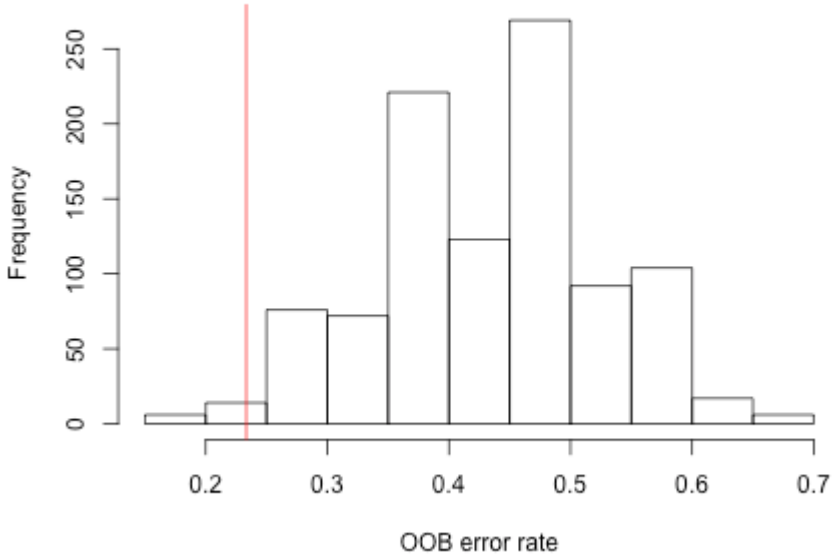


Figure 3 | OOB error rate calculated for 1000 random forest classifiers built using permuted classes

OOB error rate for the classifier built using the true classes is indicated with a red line.

Table 1 | Confusion matrix built from predictions on training data made by the random forest model

		Prediction		**Classification error**
		Gastrointestinal	Invasive	
Class	Gastrointestinal	19	1	0.05
	Invasive	6	4	0.6

Functions associated with key predictors identified using a random forest approach

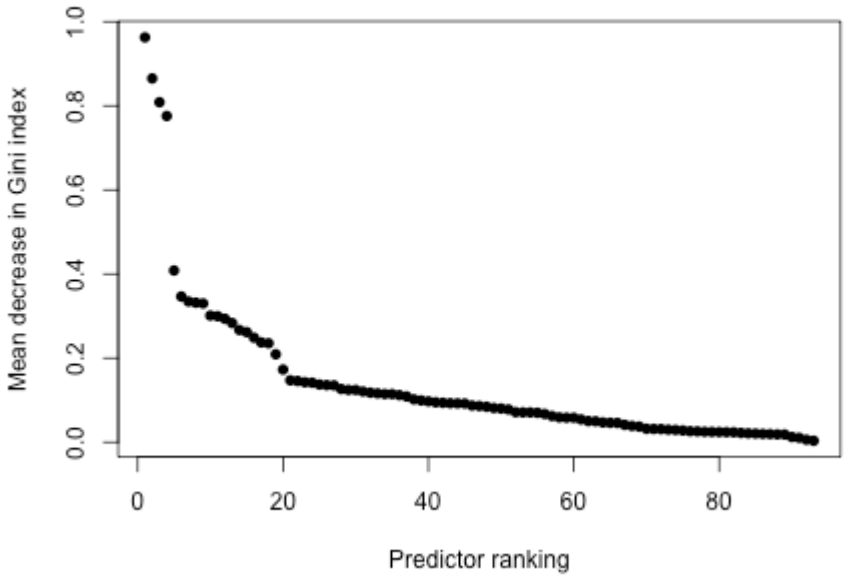


Figure 4 | Variable importance of predictors used in building the random forest model

Key predictors were selected by taking those which showed a mean decrease in Gini index (feature importance) across the random forest in the highest 10% of all feature importances recorded for the permuted classes. Gini index is a measure of the increase in purity of daughter nodes achieved by splitting the data using each variable. It gives an indication of the value of each variable in identifying isolates with our phenotype of interest. The feature importance of the top predictors in the random forest model looked markedly higher than those for the rest of the training data (Figure 4), and while the model did not have high predictive accuracy, looking at the most informative genes could give us some insight into the functions that best separate the two classes. Four genes showed mean decreases in Gini index greater than that found in 95% of permuted importance values (Table 2).

Table 2 | Genes with the best value in discriminating between invasive and gastrointestinal Campylobacter, as determined by training a random forest classifier on orthologous genes and identifying the most informative using Gini index

Reference locus ID	EggNOG model	COG	Notes	Mann-Whitney U *P*-value	Mean decrease in Gini Index	Empirical *P*-value
OXC6292_01615	ENOG410I4UI	S	FUSC family protein	0.022*	0.963	0.006*
OXC6292_01140	ENOG410I2PS	M	Gne, UDP-GlcNAc/Glc 4-epimerase	0.065	0.865	0.008*
OXC6292_00702	ENOG410I2XB	D	MreB, rod shape determining protein	0.008*	0.808	0.011*
OXC6292_01248	ENOG410I5ZU	S	Hypothetical protein	0.123	0.776	0.012*
OXC6292_00469	ENOG410I39I	P	ABC transporter, spermidine/putrescine transport	0.040*	0.408	0.059
OXC6292_00355	ENOG410I4S2	S	Peptidase C39 family, bacteriocin transport, quorum sensing	0.062	0.346	0.082
OXC6292_00090	ENOG410I4U5	O	Disulfide bond formation protein	0.202	0.335	0.088
OXC6292_00266	ENOG410I2YA	S	Pgp1 LD-carboxypeptidase	0.068	0.332	0.089
OXC6292_00105	ENOG410I2P7	S	Succinyl-CoA ligase ADP-forming subunit alpha	0.215	0.330	0.090

4/10 invasive strains show truncations in the OXC6292_01615 protein, and an additional isolate shows an alternative start site, which may result in failure of the protein to be transcribed or translated (Supplementary Figure 1). In contrast, the gene is intact in all gastrointestinal isolates. The function of this protein in *Campylobacter* is not currently known, however it contains a FUSC-like domain. Truncations in this gene gave the best discriminatory ability in both the single tree and the forest ensemble approaches. The *galE/gne* gene carries some mutations that are unique to the invasive strains. This is known to be involved in LPS biosynthesis and virulence in *C. jejuni* (Fry et al., 2000). One of the top predictors, OXC6292_01248, appeared to be a successful discriminatory gene due to inconsistent annotation of the start codon across genomes.

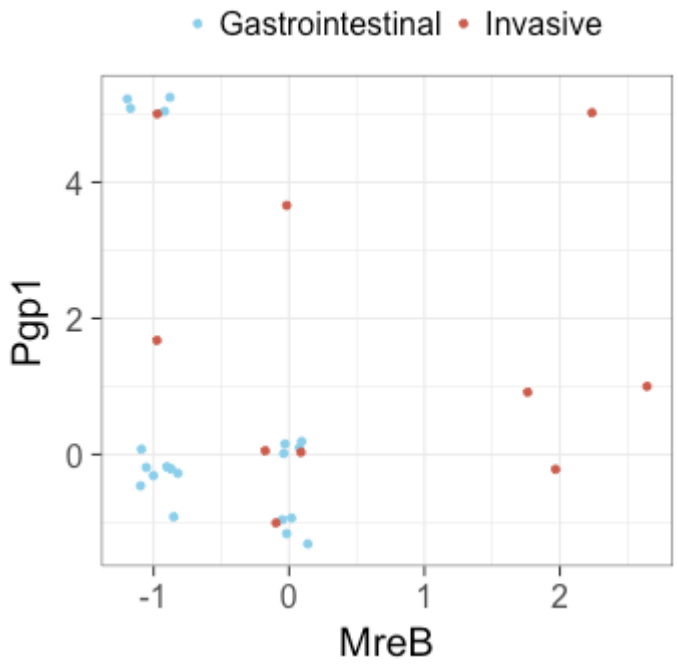


Figure 5 | DBS distributions for invasive and gastrointestinal lines for two cell shape determining proteins identified as associated with invasiveness

The position of the points is offset slightly (up to 0.2 bits) to allow discrimination of strains with identical scores for each gene.

The *mreB* gene has acquired several mutations in invasive lines that were not observed in any gastrointestinal isolates. The MreB protein has a role in maintaining cell shape and adherence to host cells, and is upregulated in response to mucins secreted by the human gut (Tu et al., 2008). Mutations in *mreB* in invasive isolates scored between 1.8 and 2.8 (empirical *P*-value 0.17-0.11), indicating subtle changes in sequence that are unlikely to impair protein function. In addition to *mreB*, *pgp1* appeared in the top predictors, suggesting that changes in cell shape may be associated with this shift in lifestyle (Figure 5, compare to Supplementary Figure 2). While most sequence differences consisted of amino acid substitutions, one invasive isolate had a mutation that would result in the truncation of the Pgp1 protein. Because *pgp1* and *pgp2* have recently both been linked with cell shape switching in *C. jejuni* (Esson et al., 2016), we investigated sequence changes in *pgp2* as well. We saw no truncations in *pgp2* in any of our invasive isolates, but one isolate carried mutations that gave it a score of 4.2 (empirical *P*-value 0.06). Overall, 7/10 invasive isolates showed deviation from modelled sequence constraints in either *mreB* or *pgp1*, compared to 4/20 gastrointestinal isolates, suggestive of a relationship between mutations in cell shape determining proteins and a transition to an invasive lifestyle.

Previously identified markers of invasive potential were of no value in this investigation

Genetic markers of invasive potential in *Campylobacter* have been identified previously, so we investigated whether any of these were informative in our dataset. A previously identified genetic marker, called “invasion associated marker” or *iam* was previously found to be associated with invasive infection (Carvalho et al., 2001). The marker was found in all of the isolates in our study, however no significant association between bitscore at this locus and invasive potential could be found (*P*-value = 0.68). Loci identified in other studies, *peb1* (Pei & Blaser, 1993), *peb4A* (Burucoa et al., 1995) and *cadF* (Konkel et al., 1997) were found in these isolates but also showed so significant association with invasiveness (*P* = 0.73, 0.84 and 1.00, respectively), all owing to low sequence diversity in these genes. This suggests that the isolates in our study all had genetic backgrounds that could have facilitated invasive infection, but only our New Zealand isolates achieved this outcome.

Discussion

This study aimed to address whether there was a common signature associated with adaptation to an invasive lifestyle in a sampling of invasive *Campylobacter jejuni* found in clinical settings across New Zealand. If there is a single path to invasiveness in *Campylobacter*, we might expect this to be reflected in the genomic data, however if there are multiple independent routes to achieving an invasive phenotype, this signal is likely to be difficult to identify in such a small sampling of bacteria.

No protein coding genes investigated showed consistent signs of functional degradation unique to invasive isolates. This was to be expected, due to the close relationships between individual invasive strains and associated reference gastrointestinal strains, compared to the relationships between the invasive strains themselves. For a consistent signal to be present, mutations in the same gene would have to occur in all invasive isolates. It may be that it takes longer for deleterious mutations to accumulate in these invasive strains than the time that elapsed before sequencing of these isolates, and for uninformative and mildly deleterious changes that have occurred by chance to be purged from these genomes by negative selection. All strains in the study carried previously identified genetic markers of invasive potential, suggesting that either these are not sufficient for invasive infection, or that host and environmental factors encountered by these New Zealand isolates were more permissive of invasive infections.

Because our collection of invasive isolates was so small, caution needs to be taken in the interpretation of the findings of this study. Separation of invasive and gastrointestinal isolates using a recursive partitioning technique indicated that a subset of invasive isolates could be identified using a single discriminatory gene, but that following this initial split, little additional discriminatory power was added by successive variables. This indicates that the first node captured most of the differences between the two classes, and each successive node identified weaker trends that may have been alternative signatures of invasiveness, or could have been unique aspects of the isolates in our study that are unlikely to appear in an independent set of invasive strains. These patterns may replicate in larger datasets, and there are certainly biological explanations that can be drawn for why some of the genes identified show stronger associations with invasiveness than others, but with a study size this small we also face a potentially high burden of false positive results.

One may expect combinations of genes with accurate discriminatory power for such a small sampling of bacteria to arise at random, as we saw here with a subset of recursive partitioning trees built to classify strains based on randomly assigned labels, where five or six genes were usually sufficient to discriminate strains perfectly. Permutation testing indicates that the associations between adherence to sequence constraints and phenotype we identified in these invasive strains are stronger than the majority of those identified using randomised data. A random forest based approach gave similar predictive power, however the out-of-bag classifier correctly identified less than half of the invasive strains, indicating that the data are not sufficient to confidently discriminate the clinical outcomes caused by these bacteria using only orthologous gene sequences. In this case, host factors such as age, gender and immune-compromising conditions (Louwen et al., 2012) may have played a significant role in the differences in infection outcomes. Independent events each affecting invasiveness may have occurred in each strain, in which case the study will be unable to detect these. Changes in noncoding sequence also could have impacted invasive potential, for example by changing gene expression patterns (Bailey et al., 2014).

Some of the genes that have been identified as top predictors of invasive infection are supported by the existing literature, particularly those that show truncations in invasive strains but not in gastrointestinal strains. We discovered that mutations in two genes that potentially control cell shape considered collectively were effective at separating a subset of invasive and gastrointestinal strains. Cell shape has been tied to pathogenicity in the past,

with the helical shape of *Campylobacter* cells specifically implicated in colonisation ability (Doble et al., 2012; Ferrero & Lee, 1988; Frirdich et al., 2012, 2014). The mutations we have observed in MreB appear to be substitutions which are unlikely to completely disrupt protein function. Some long term evolution experiments in *E. coli* have identified convergent changes in MreB relating to adaptation to new conditions, such as higher temperatures (Tenaillon et al., 2012), and minimal media (Deatherage et al., 2014). Interaction of other protein components with MreB has been found to alter pathogenicity traits, for example disruption of MreC and MreD cause downregulation of SPI-1 in Salmonella, which is essential for invasion and motility (Doble et al., 2012). MreB has also been implicated in mediating the positioning of cell-wall associated virulence-related proteins (Cowles & Gitai, 2010; Mauriello et al., 2010). Some cells are known to transition to coccid shapes under conditions of stress. In *Vibrio parahaemolyticus*, this process is accompanied by changes in expression and localisation of MreB (Chiu et al., 2008). It may be that we are seeing slight changes to the way MreB is arranged or interacts with other proteins as a response to the change in environment experienced by *Campylobacter* transitioning to an invasive lifestyle.

The Pgp1 protein was also identified as informative during this analysis. Switching of *C. jejuni* from helical to rod-shaped has been found to occur repeatedly in standard laboratory conditions due to phase variation resulting from the variation in an 8-A tract within the gene (Esson et al., 2016). In the Esson et al. study, all isolates that showed variation in the length of this tract had switched from helical to rod-shaped, suggesting that our isolate with a truncation in this protein has likely also switched to a rod-shaped phenotype. Deletion of *pgp1* has been shown to impair chick colonisation (Frirdich et al., 2012), suggesting that disruptive mutations in the gene may impair gastrointestinal colonisation in humans. If this is the case then these mutations may have occurred after invasive infection occurred, and could represent a dead end in the transmission cycle of these pathogens.

In addition to cell shape determining genes, the UDP-GlcNAc/Glc 4-epimerase Gne also showed more mutations in invasive strains than in gastrointestinal close relatives. The Gne gene provides Gal and GalNAc for the major cell-surface carbohydrates: lipopolysaccharides, capsule, and glycoprotein N-linked heptasaccharide (Bernatchez et al., 2005). Gne has also been shown to play a role in adhesion and invasion of host cells in *C. jejuni*, with mutants showing severe impairment (Fry et al., 2000). Alterations in the Gne protein could again signify relaxed selection on this protein in the new environment, adaptive

change, or a deleterious mutation that was not eliminated from the population due to a population bottleneck that occurred during infection.

Concluding statements

While some of the key indicator genes identified in this study look promising, we advise caution in interpreting the results of this study, as we lacked statistical power to assign a high degree of confidence to our findings. The dataset used in this study is not of a size typically used to build a supervised classifier using machine learning techniques, however we anticipated that using this approach would provide a more nuanced measure of the discriminatory power of the genes in the study than classic statistical techniques, as random forests have been found to be a particularly effective way of dealing with sparse data (Bowles, 2015), and are being explored as an effective tool for finding associations between genetic variants and phenotypes of interest for this reason (Goldstein et al., 2010; Touw et al., 2013). While the performance of the predictor was better than most models trained using permuted classes, the likelihood that the model would be predictive in a separate set of invasive isolates is low, due to the small number of informative genes identified, the limited classification accuracy for the invasive class and the large potential for overtraining in this case. The findings of this study on their own should be taken with skepticism, but given support in the literature for an association between cell shape and cell surface polysaccharides, they may be worth investigating further in larger datasets. The workflow used to perform this analysis is straightforward and requires little manual intervention to perform, therefore similar investigations could be repeated with other collections of invasive pathogens, allowing for cross-validation of results.

References

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Retrieved from http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Aronesty, E. (2011). ea-utils: Command-line tools for processing biological sequencing data. *Expression Analysis, Durham, NC*.

Bailey, S. F., Hinz, A., & Kassen, R. (2014). Adaptive synonymous mutations in an experimentally evolved Pseudomonas fluorescens population. *Nature Communications*, *5*, 4076.

Bernatchez, S., Szymanski, C. M., Ishiyama, N., Li, J., Jarrell, H. C., Lau, P. C., … Wakarchuk, W. W. (2005). A single bifunctional UDP-GlcNAc/Glc 4-epimerase supports the synthesis of three cell surface glycoconjugates in Campylobacter jejuni. *The Journal of Biological Chemistry*, *280*(6), 4792–4802.

Bowles, M. (2015). *Machine Learning in Python: Essential Techniques for Predictive Analysis*. John Wiley & Sons.

Breiman, L. (1996). *Out-of-bag estimation*. Statistics Department, University of California Berkeley, Berkeley CA 94708.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees*. Chapman and Hall/CRC.

Burucoa, C., Frémaux, C., Pei, Z., Tummuru, M., Blaser, M. J., Cenatiempo, Y., & Fauchère, J. L. (1995). Nucleotide sequence and characterization of peb4A encoding an antigenic protein in Campylobacter jejuni. *Research in Microbiology*, *146*(6), 467–476.

Calva, J. (1988). Cohort study of intestinal infection with Campylobacter in Mexican children. *The Lancet*, *331*(8584), 503–506.
Carvalho, A. C. T., Ruiz-Palacios, G. M., Ramos-Cervantes, P., L.-E., C., Jiang, X., & Pickering, L. K. (2001). Molecular Characterization of Invasive and Noninvasive Campylobacter jejuni and Campylobacter coli Isolates. *Journal of Clinical Microbiology*, *39*(4), 1353–1359.
Chiu, S.-W., Chen, S.-Y., & Wong, H.-C. (2008). Localization and Expression of MreB in Vibrio parahaemolyticus under Different Stresses. *Applied and Environmental Microbiology*, *74*(22), 7016–7022.
Cody, A. J., McCarthy, N. M., Wimalarathna, H. L., Colles, F. M., Clark, L., Bowler, I. C. J. W., … Dingle, K. E. (2012). A longitudinal 6-year study of the molecular epidemiology of clinical campylobacter isolates in Oxfordshire, United kingdom. *Journal of Clinical Microbiology*, *50*(10), 3193–3201.
Cowles, K. N., & Gitai, Z. (2010). Surface association and the MreB cytoskeleton regulate pilus production, localization and function in Pseudomonas aeruginosa. *Molecular Microbiology*, *76*(6), 1411–1426.
Cox, M. P., Peterson, D. A., & Biggs, P. J. (2010). SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics*, *11*, 485.
Deatherage, D. E., Traverse, C. C., Wolf, L. N., & Barrick, J. E. (2014). Detecting rare structural variation in evolving microbial populations from new sequence junctions using breseq. *Frontiers in Genetics*, *5*, 468.
Doble, A. C., Bulmer, D. M., Kharraz, L., Karavolos, M. H., & Khan, C. M. A. (2012). The function of the bacterial cytoskeleton in Salmonella pathogenesis. *Virulence*, *3*(5), 446–449.
Esson, D., Mather, A. E., Scanlan, E., Gupta, S., de Vries, S. P. W., Bailey, D., … Grant, A. J. (2016). Genomic variations leading to alterations in cell morphology of Campylobacter spp. *Scientific Reports*, *6*, 38303.
Fauchere, J. L., Rosenau, A., Veron, M., Moyen, E. N., Richard, S., & Pfister, A. (1986). Association with HeLa cells of Campylobacter jejuni and Campylobacter coli isolated from human feces. *Infection and Immunity*, *54*(2), 283–287.
Ferrero, R. L., & Lee, A. (1988). Motility of Campylobacter jejuni in a viscous environment: comparison with conventional rod-shaped bacteria. *Journal of General Microbiology*, *134*(1), 53–59.
Frirdich, E., Biboy, J., Adams, C., Lee, J., Ellermeier, J., Gielda, L. D., … Gaynor, E. C. (2012). Peptidoglycan-modifying enzyme Pgp1 is required for helical cell shape and pathogenicity traits in Campylobacter jejuni. *PLoS Pathogens*, *8*(3), e1002602.
Frirdich, E., Vermeulen, J., Biboy, J., Soares, F., Taveirne, M. E., Johnson, J. G., … Gaynor, E. C. (2014). Peptidoglycan LD-carboxypeptidase Pgp2 influences Campylobacter jejuni helical cell shape and pathogenic properties and provides the substrate for the DL-carboxypeptidase Pgp1. *The Journal of Biological Chemistry*, *289*(12), 8007–8018.
Fry, B. N., Feng, S., Chen, Y. Y., Newell, D. G., Coloe, P. J., & Korolik, V. (2000). The galE gene of Campylobacter jejuni is involved in lipopolysaccharide synthesis and virulence. *Infection and Immunity*, *68*(5), 2594–2601.
Gastro bug hit 5000. (2016, August 30). Gastro bug hit 5000 in Havelock North. *Radio New Zealand*. Retrieved from http://www.radionz.co.nz/news/national/312100/gastro-bug-hit-5000-in-havelock-north
Goldstein, B. A., Hubbard, A. E., Cutler, A., & Barcellos, L. F. (2010). An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings. *BMC Genetics*, *11*, 49.
Goossens, H., Henocque, G., Kremp, L., Rocque, J., Boury, R., Alanio, G., … Macart, M. (1986). Nosocomial outbreak of Campylobacter jejuni meningitis in newborn infants. *The Lancet*, *2*(8499), 146–149.
Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., … Bork, P. (2016). eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research*, *44*(D1), D286–93.
Huson, D. H., & Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, *23*(2), 254–267.
Huynh-Thu, V. A., Wehenkel, L., & Geurts, P. (2008). Exploiting tree-based variable importances to selectively identify relevant variables. *Proceedings of FSDM08, ECML/PKDD Workshop on New challenges for feature selection in data mining and knowledge discovery* (pp. 60–73).
Kirk, M. D., Pires, S. M., Black, R. E., Caipo, M., Crump, J. A., Devleesschauwer, B., … Angulo, F. J. (2015). World Health Organization Estimates of the Global and Regional Disease Burden of 22 Foodborne Bacterial, Protozoal, and Viral Diseases, 2010: A Data Synthesis. *PLoS Medicine*, *12*(12), e1001921.
Konkel, M. E., Garvis, S. G., Tipton, S. L., Anderson, D. E., Jr, & Cieplak, W., Jr. (1997). Identification and molecular cloning of a gene encoding a fibronectin-binding protein (CadF) from Campylobacter jejuni. *Molecular Microbiology*, *24*(5), 953–963.
Lake, R., Hudson, A., Cressey, P., & Gilbert, S. (2007). *Risk Profile: Campylobacter jejuni/coli in poultry (whole and pieces)*. Institute of Environmental Science & Research Limited. Retrieved from http://www.foodsafety.govt.nz/elibrary/industry/Risk_Profile_Campylobacter_Jejuni-Science_Research.pdf
Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359.
Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22.
Li, L. (2003). OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Research*, *13*(9), 2178–2189.
Louwen, R., Rogier, L., van Baarlen, P., van Vliet, A. H. M., van Belkum, A., Hays, J. P., & Endtz, H. P. (2012). Campylobacter bacteremia: A rare and under-reported event? *European Journal of Microbiology & Immunology*, *2*(1), 76–87.

Mauriello, E. M. F., Mouhamar, F., Nan, B., Ducret, A., Dai, D., Zusman, D. R., & Mignot, T. (2010). Bacterial motility complexes require the actin-like protein, MreB and the Ras homologue, MglA. *The EMBO Journal*, *29*(2), 315–326.

Pappu, V., & Pardalos, P. M. (2014). High-Dimensional Data Classification. In F. Aleskerov, B. Goldengorin, & P. M. Pardalos (Eds.), *Clusters, Orders, and Trees: Methods and Applications* (pp. 119–150). Springer New York.

Pei, Z., & Blaser, M. J. (1993). PEB1, the major cell-binding factor of Campylobacter jejuni, is a homolog of the binding component in gram-negative nutrient transport systems. *The Journal of Biological Chemistry*, *268*(25), 18717–18725.

Ruiz-Palacios, G. M., Torres, J., Torres, N. I., Escamilla, E., Ruiz-Palacios, B. R., & Tamayo, J. (1983). Cholera-like enterotoxin produced by Campylobacter jejuni. Characterisation and clinical significance. *The Lancet*, *2*(8344), 250–253.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* , *30*(14), 2068–2069.

Skirrow, M. B., Jones, D. M., Sutcliffe, E., & Benjamin, J. (1993). Campylobacter bacteraemia in England and Wales, 1981-91. *Epidemiology and Infection*, *110*(3), 567–573.

Tenaillon, O., Rodríguez-Verdugo, A., Gaut, R. L., McDonald, P., Bennett, A. F., Long, A. D., & Gaut, B. S. (2012). The molecular diversity of adaptive convergence. *Science*, *335*(6067), 457–461.

Therneau, T., Atkinson, B., & Ripley, B. (2015). rpart: Recursive Partitioning and Regression Trees (Version 4.1-10). Retrieved from https://CRAN.R-project.org/package=rpart

Touw, W. G., Bayjanov, J. R., Overmars, L., Backus, L., Boekhorst, J., Wels, M., & van Hijum, S. A. F. T. (2013). Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? *Briefings in Bioinformatics*, *14*(3), 315–326.

Tu, Q. V., McGuckin, M. A., & Mendz, G. L. (2008). Campylobacter jejuni response to human mucin MUC2: modulation of colonization and pathogenicity determinants. *Journal of Medical Microbiology*, *57*(Pt 7), 795–802.

Wheeler, N. E., Barquist, L., Kingsley, R. A., & Gardner, P. P. (2016). A profile-based method for identifying functional divergence of orthologous genes in bacterial genomes. *Bioinformatics* , *32*(23), 3566–3574.

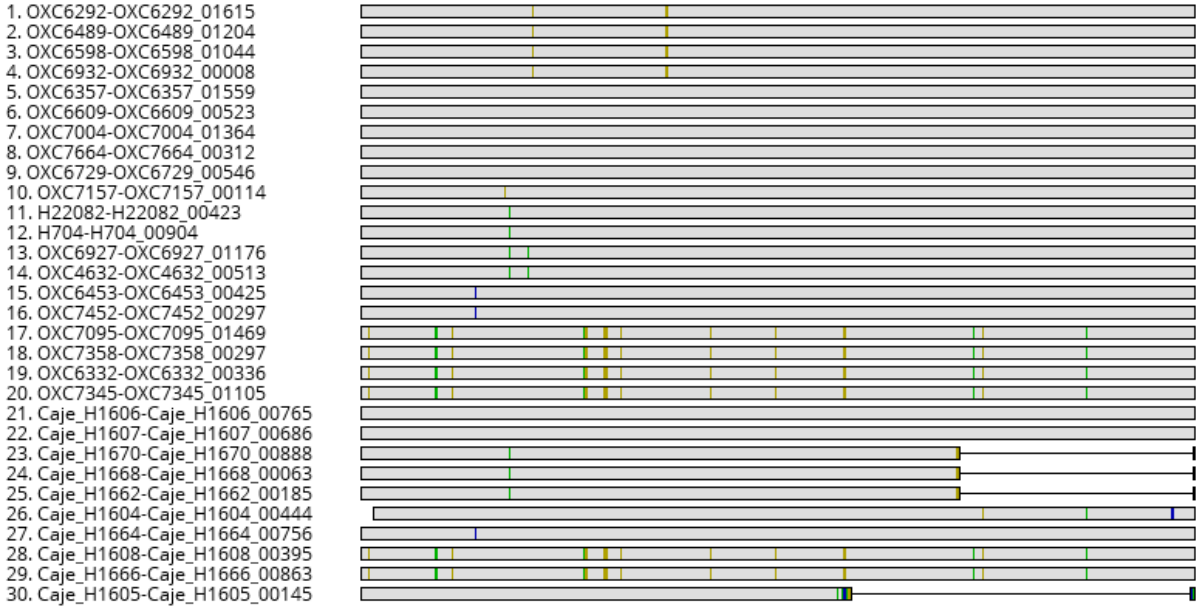
Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–829.

Supplementary Material

Supplementary Table 1 | Summary statistics for genes that show the most extreme differences in bitscores between invasive and gastrointestinal samples

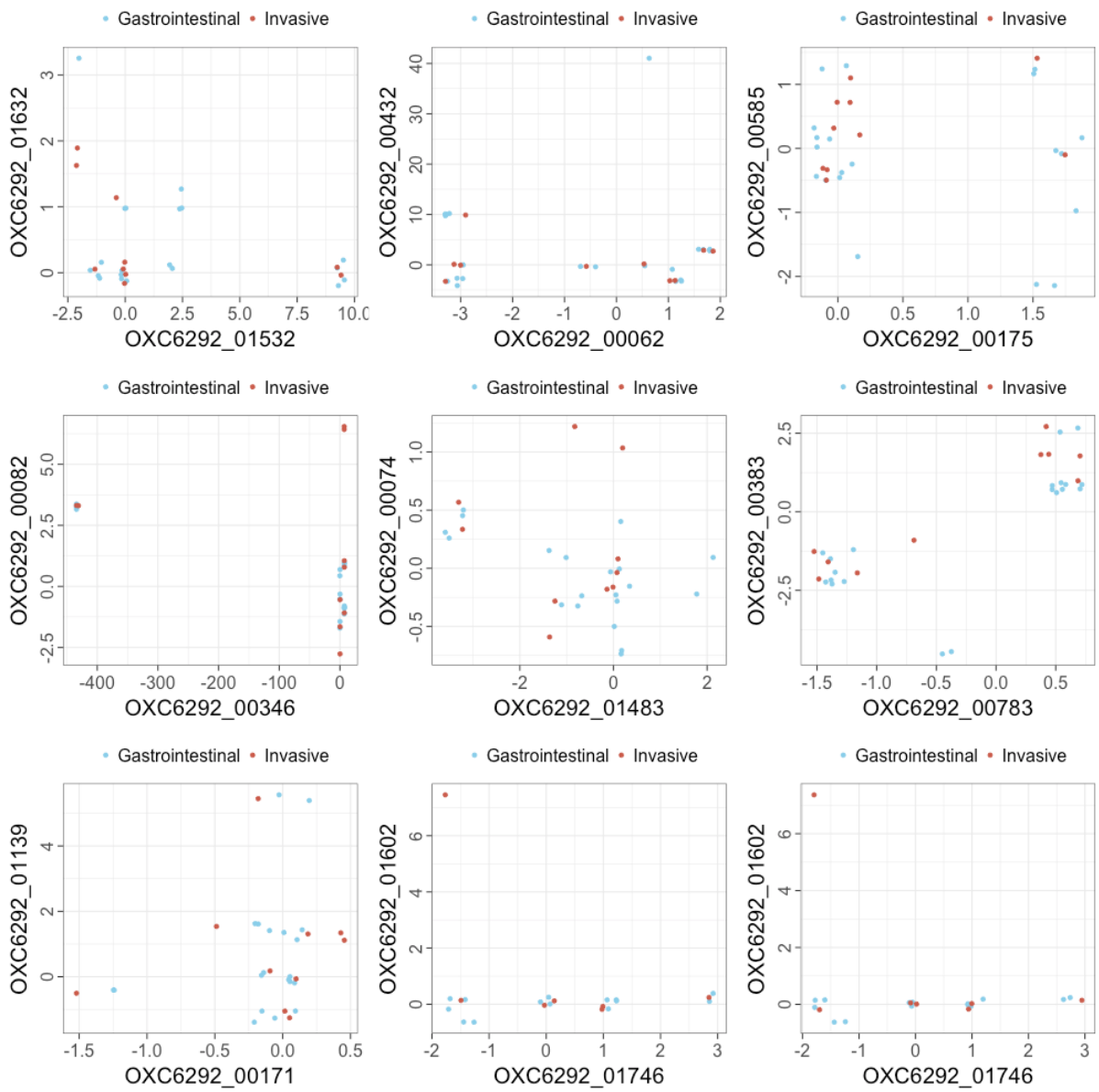
Comments based on a visual inspection of sequence alignments are provided as an assessment of whether these are likely to be true differences, or differences in annotation methods between the genome sets.

Gene	Difference in median bitscore	Bitscore empirical p-value	KS statistic	KS empirical p-value	Possibility of misannotation contributing to differences
OXC6292_01248	-1.95	0.033	0.400	0.005	Y
OXC6292_01147	-2.8	0.022	0.350	0.015	N
OXC6292_01615	6.5	0.012	0.500	0.001	N
OXC6292_00095	18.3	0.006	0.450	0.003	Y
OXC6292_00639	-2	0.031	0.350	0.015	Y
OXC6292_00465	-5.5	0.013	0.329	0.020	Y
OXC6292_00354	190.45	0.000	0.438	0.004	N
OXC6292_00355	7.2	0.011	0.563	0.000	Y
OXC6292_01685	2.65	0.024	0.500	0.001	N
OXC6292_00886	-1.8	0.036	0.381	0.014	Y
OXC6292_00690	-51.6	0.003	0.333	0.019	Y



Supplementary Figure 1 | Sequence alignment of OCX6292_01615 homologs

Truncations can be seen in 4/10 invasive isolates, as well as a change in start position for an additional isolate which may result in failure to transcribe the protein.



Supplementary Figure 2 | Score distributions for random pairings of genes

This figure gives an indication of the deviation of bitscores of invasive isolates from their gastrointestinal close relatives for genes not identified as predictive of invasive infection.

Chapter Five | Draft: Profile-based analysis of *Salmonella* reveals signatures of invasiveness

Preface

While the study performed in Chapter Four identified interesting candidate genes that may have been involved in adaptation to an invasive phenotype, our power to draw broader conclusions about the adaptation of other *Campylobacter* to an invasive lifestyle was limited. A better opportunity for identifying broader trends presented itself in a revisiting of the study by Nuccio and Bäumler (2014). The invasive *Salmonella* used in this study had transitioned to an invasive lifestyle a long time ago, so a larger amount of mutations are expected to have accumulated since the change in lifestyle, and more time will have passed for selection to act on these mutations. Thus, we expect mutations shared by these isolates to be more characteristic of changes that are adaptive or neutral after a shift to an invasive lifestyle.

The investigation performed in this chapter is an extension of work initially performed in Chapter Two. In Chapter Two, data from Nuccio and Bäumler (2014) were used to test the pairwise comparative DBS method for detecting genomic changes associated with host adaptation in *Salmonella*. In this study, I wanted to focus on invasiveness, a phenotype which is often closely linked with host adaptation in *Salmonella*, but is likely tied to more consistent selective pressures than adaptation to a range of different hosts.

This chapter again uses the random forest approach trialed in the previous chapter, and demonstrates that trends identified using this method mirror genetic changes associated with invasiveness identified in previous studies using pseudogene data.

Contributions

I performed the analysis, Lars Barquist and Paul Gardner helped in the planning of the project and gave feedback on the manuscript.

Profile-based analysis of *Salmonella* genomes reveals signatures of invasiveness

Nicole E. Wheeler[1], Lars Barquist[2], Paul P. Gardner[1]

1. School of Biological Sciences, University of Canterbury, Christchurch, New Zealand.
2. Institute for Molecular Infection Biology, University of Wuerzburg, Wuerzburg, Germany.

Abstract

Salmonella enterica is a pathogen of global importance. Members of the species fall along a continuum, from pathovars which cause gastrointestinal infection and low mortality, to those that cause invasive infection and high mortality. These two variants of *Salmonella* have been studied extensively in the past, however many investigations have involved either generalised insights gained from large numbers of genomes, or information on changes in gene content and degradation based on a smaller number of genomes. Here we present an approach for gaining broader insight into the differences in gene degradation between invasive and gastrointestinal pathovars of *Salmonella*. We find that this approach identifies patterns captured by more labour-intensive investigations, but can be readily scaled to larger analyses.

Introduction

Salmonella enterica is a pathogen of global importance, with an estimated 93.8 million cases of gastroenteritis and 155,000 deaths occurring due to *Salmonella* per year (Majowicz et al., 2010). *Salmonella enterica* serovars typically lie along a spectrum that includes those that have a broad host range and cause self-limiting gastrointestinal infection, and those that are more restricted in host range, but cause more systemic disease and are typically associated with higher mortality (Rabsch et al., 2002). These host-restricted, invasive variants of *Salmonella enterica* have evolved independently several times from gastrointestinal ancestors (Klemm et al., 2016), giving us an opportunity to look for parallel genomic changes that have accumulated in invasive but not in gastrointestinal pathovars, which are indicative of the differences in selective pressures applied by these two lifestyles.

Accumulation of pseudogenes has been found to be a common feature of host-adapted *Salmonella* (Holt et al., 2009). Studies have found common patterns in the functions of genes that become pseudogenes in invasive *Salmonella* (Langridge et al., 2015; Nuccio &

Baumler, 2014), indicating that these loss of function mutations could serve as a good tool for identifying novel invasive strains, and could help us understand the different functional requirements of the two niches.

In this study, we have used a recently developed method for detecting divergence of protein sequences from modelled sequence constraints (Wheeler et al., 2016) to identify protein coding genes that show greater divergence from sequence constraints associated with either the invasive or the gastrointestinal niche. In contrast with other studies that have predominantly looked for more definitive signs of pseudogenization, such as frameshifts and premature stop codons, our approach also looks for more subtle deviations from predicted sequence constraints.

We extend the approach used in (Wheeler et al., 2016), by incorporating a machine-learning based approach to identify the most useful genes for distinguishing the two lifestyles. Random forests are a popular machine-learning technique that can handle large datasets with a low ratio of informative to uninformative variables, and deal with complex interactions between variables such as epistatic interactions (Pappu & Pardalos, 2014; Touw et al., 2013). They work by building an ensemble of classification and regression trees designed to predict a characteristic of the samples, in this case invasiveness. The process of building a random forest will produce measures of variable importance that can be used to assess the relative utility of different genes in classification of *Salmonella* strains based on lifestyle.

Using this approach we have identified a shortlist of genes that appear to be most informative in separating invasive and gastrointestinal isolates. Our findings reflect those of other studies (Holt et al., 2009; Langridge et al., 2015; Nuccio & Baumler, 2014), in identifying a common set of metabolic pathways that appear to be degraded in invasive isolates. We demonstrate that this approach can easily integrate information on the consequences of different SNPs within the same gene into a comparable, quantitative measure that can be used to identify associations between genetic variation and phenotypic traits.

Methods

Genome data and identification of orthologs

Genomes for 13 *Salmonella enterica* strains were retrieved from the NCBI database (accessions and strain information can be found in Supplementary Table 1). The strains

were divided into gastrointestinal and invasive serovars according to the classifications made by Nuccio and Bäumler (2014). Ortholog calls were also taken from the Supplementary Material of Nuccio and Bäumler (2014).

Measuring the divergence of genes from predicted sequence constraints

Profile hidden Markov models (HMMs) for gamma-proteobacterial proteins were retrieved from the eggNOG database (Huerta-Cepas et al., 2016). Each protein sequence was searched against the HMM database using hmmsearch from the HMMER3.0 package (http://hmmer.org). The top scoring model corresponding to each protein was used for analysis (N = 8060 groups). Orthologous groups with no model hit, or more than one top model hit were excluded from analysis (N = 1524, 6536 remained). Genes within remaining orthologous groups with no significant hit to a model were assigned a score of zero, reflecting a loss of the function of that protein. Additionally, genes with no variation in bitscore for the match between protein sequences and their respective eggNOG HMM across isolates were excluded (N = 188, 6438 remained).

Training a random forest classifier

The R package “randomForest” (Liaw & Wiener, 2002) was used to build random forest classifiers using a variety of parameters to assess which were best for accuracy. Prediction accuracy, as measured by out-of-bag (OOB) error rate, stabilised at 1000 trees, so we chose this as a parameter for optimising the number of genes sampled per node (mtry). mtry values of 1, $p/10$, $p/5$, $p/3$, $p/2$ and p (where p = the number of predictors) were tested, and we found that at mtry=$p/10$, the number of genes that were either not incorporated into trees, or did not improve the homogeneity of daughter nodes when they were incorporated into trees (as measured by mean decrease in Gini index, Breiman et al., 1984) stabilised at ~92%.

To improve the performance of the model, we performed five model building and sparsity pruning cycles. For the first cycle, we built a model using all genes that met the inclusion criteria, and performed sparsity pruning by eliminating all variables that had a mean Gini index (variable importance) of zero or lower (meaning the gene was either not included in the model or did not improve model accuracy when it was). Four successive rounds of model building and sparsity pruning involved building a new model with the pruned dataset, then pruning the genes with the lowest 50% of variable importances. The resulting model had 100% out-of-bag classification accuracy.

Testing random forest top variables with univariate statistical analyses

Each predictor included in the initial set of 6,438 genes was tested for association with invasiveness using a linear regression, with invasiveness coded as a binary variable, and for significant differences in score between the two groups using a Kolmogorov Smirnov test and a Mann-Whitney U test. Correction for multiple testing was performed using the Benjamini Hochberg method.

Building a control model

To test whether the predictive ability of the model was due to true signal in the data, we ran the same training procedure again with randomised invasiveness classifications over five rounds of sparsity pruning. The final model had a ~23% error rate, with many strains receiving votes between 40 and 50% for both categories, indicating that confidence in predictions was low and there was high disagreement between individual decision trees.

Results

We have implemented a recently developed method for detecting deleterious mutations in protein coding genes, paired with random forests to identify informative genes for distinguishing invasive and gastrointestinal *Salmonella* pathovars. Random forests are a popular machine learning method for identifying informative variables that can deal with sparse, high-dimensional data (Pappu & Pardalos, 2014), which is common in many genome-wide association studies. They have the added benefit of being more easily interpretable than other machine learning based methods, providing measures of the relative contributions of different variables to classifier performance. Each random forest consists of an ensemble of classification and regression trees (Breiman et al., 1984), each built using a subset of cases (in this case, serovars). For each node in the tree, a subset of the predictor variables is selected, and the best variable of the set is assigned to that node in order to separate the classes (in the case of a classification problem). First, we built a model using all orthologous groups that met our filtering criteria, to get an indication of the proportion of genes providing informative signal. We then performed iterative feature selection to select the best predictor variables for our model until we achieved perfect out-of-bag model performance. We were able to show that older invasive strains are detected with greater confidence, and suggest that the voting strategy implemented by random forests, or the number of features included in the model may require adjustment for the detection of newly emerging invasive isolates.

A random forest built using all orthologous groups reveals the sparsity of the data

When the full set of orthologous genes was used to build a model, a subset of genes ranked much higher than the others in variable importance (VI) (Figure 1). We then saw a tailing off of VI, resulting in 4,721 orthologous groups either not being used in the model, or not improving classification accuracy (as indicated by VI ≤ 0). There is an element of stochasticity in VI measures for random forest models. When the same model was built 10 times using the same dataset and parameters, 633 genes showed VI>0 in all models, and another 2,864 genes showed VI>0 in at least one model.

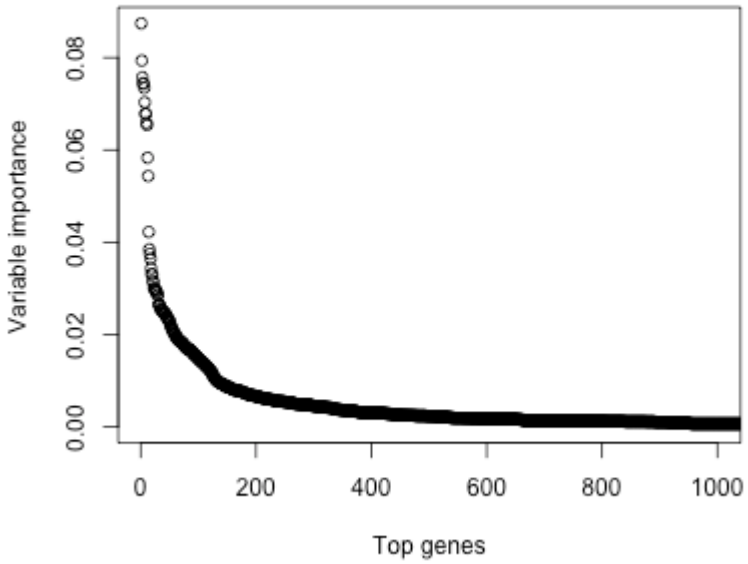


Figure 1 | Variable importance of the top 1000 genes used in the original training set Variable importance was measured as average decrease in Gini index in a random forest model trained on all orthologous groups that met the inclusion criteria (N = 6,438).

Further feature selection produces a perfect predictor

Repeated rounds of selecting the top 50% of predictors and re-constructing the model eventually resulted in a model with perfect performance. The final model used 196 of the original 6,438 genes for prediction. The structure of the trees in the final forest appears to be remarkably simple. Most trees consist of a single gene. This indicates that for the most part, individual trees are checking for degradation of a single gene that is frequently degraded in invasive pathovars, then votes from each tree are collated to make a decision on invasiveness. As a result, a pathovar needs to have shown degradation in at least half of the

genes in the forest to be classified as invasive. Some genes will be weighted more than others in the forest because they are lost more consistently, and therefore may represent adaptive losses more than stochastic losses due to reduced selection.

Predictive genes are typically highly correlated with invasiveness

Many of the genes that were selected for the final model had significant associations with invasiveness when tested individually (N = 32, adjusted P-value < 0.05, linear regression). However, many of the variables did not show even a nominally significant association with phenotype. These were predominantly cases where invasive isolates had lower bitscores than most gastrointestinal isolates, but a single outlier from the gastrointestinal isolates obscured the association. 106/197 genes showed significant differences in score distribution between gastrointestinal and invasive isolates, as measured by a Kolmogorov-Smirnov test (adjusted P-value < 0.05), and 163/198 showed higher values in one group compared to the other (Mann-Whitney U test, adjusted P-value < 0.05). In general, more predictors were significant when tested using a Mann-Whitney U test rather than a linear regression, which can be explained by the lack of robustness of linear regressions to outliers, which were common in this analysis. In contrast, if *P*-values for all tests are corrected for multiple testing using the total number of genes originally included in the study, 6,438, only three of the associations remain significant, and none of the tests for a shift in score distribution are significant, indicating that none of these patterns would have been detected using univariate statistical approaches. This result indicates that a strength of random forest predictors is detecting changes that occur in only a subset of isolates in the study, but are nevertheless predictive of phenotype when combined with other predictors. This is likely to be the case in newly emerging invasive isolates, where relaxed selection on a pathway will eventually lead to stochastic loss of function of genes in the pathway, but not necessarily the same genes in each isolate.

S. Dublin and S. Enteritidis pathovars are more difficult to classify than others

During model building, accuracy was initially high, but a great deal of sparsity pruning was required to correctly identify *S.* Dublin as invasive and *S.* Enteritidis as gastrointestinal. This is reflective of their relatively recent divergence and niche adaptation compared to other strains in the study. Interestingly *S.* Gallinarum was identified much more readily, however it was also misidentified as gastrointestinal by the first model, that was built using all predictors.

S. Enteritidis was mis-classified as invasive, indicating that it shared many of the genomic trends identified in invasive lineages. Genomic analyses indicate it previously carried an intact SPI that would have likely conferred invasive ability (Langridge et al., 2015), suggesting that perhaps its ancestor inhabited this niche for a time. This could explain the greater number of disrupted and deleted genes relative to other gastrointestinal strains used in this study (Nuccio & Baumler, 2014), and the difficulty in classifying it correctly. Conversely, *S*. Dublin was mis-classified as gastrointestinal. *S.* Dublin showed characteristics of genome degradation more characteristic of gastrointestinal isolates than of invasive isolates in a study by Nuccio and Bäumler (2014), and fewer pseudogenes in a study by Langridge et al. (2015), indicating that there may be fewer signs of adaptation to an invasive lifestyle in *S*. Dublin than in other invasive pathovars. *S*. Dublin appears to be less host restricted than the other invasive serovars, which could help to explain this result (Betancor et al., 2012).

We can see from the distribution of votes across successive models (Figure 2) that the correct OOB classifications are only achieved by the fifth iteration of the model. This may indicate that even greater sparsity pruning would be required to detect more recent emergence of invasive isolates, or it may indicate that recently emerged isolates cannot be distinguished due to the slow rate of accumulation of these deleterious mutations. The distribution of VI values for different predictor genes, along with the score distributions of the top predictors indicates that a small number of genes show far more discrete score distributions between the two classes of pathovar, suggesting that some losses may occur more consistently due to selective pressures.

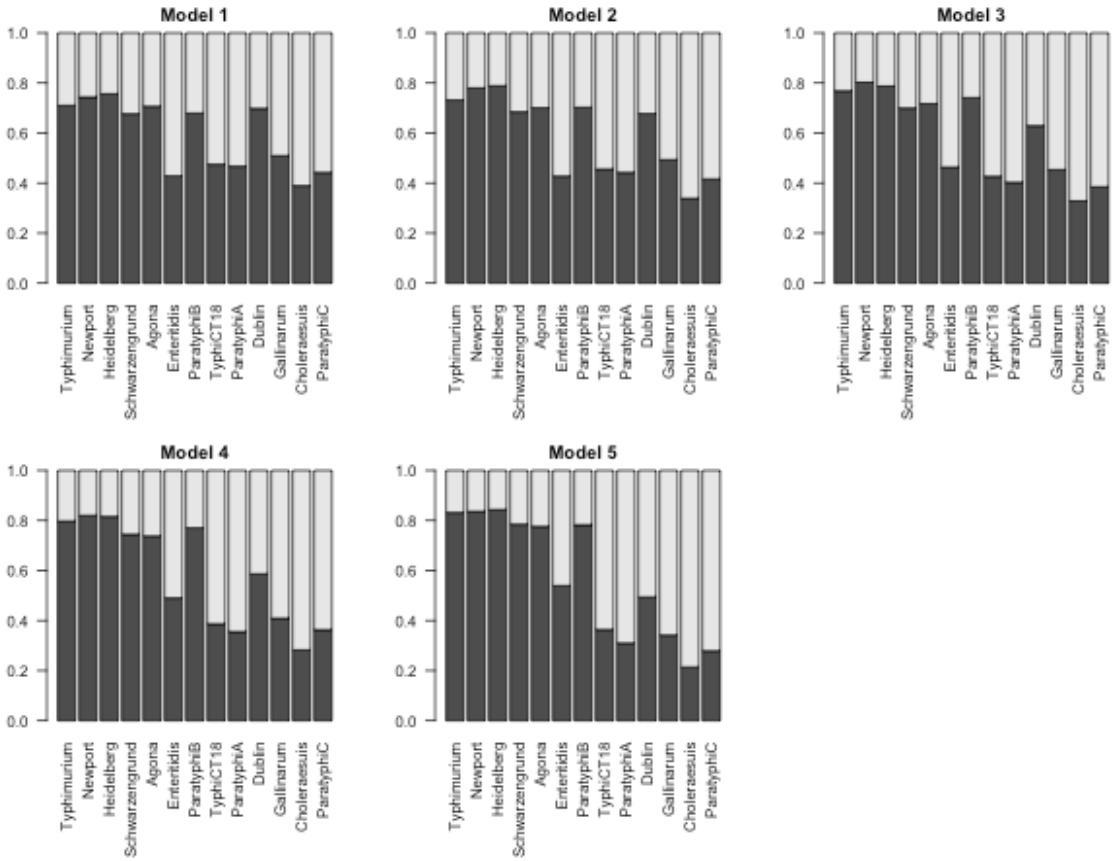


Figure 2 | Out-of-bag votes of pathovar classification for each model

Model 1 is the model built using all predictor variables, then each successive model was built using sparsity pruning from the previous model's predictor variables. Model 5 is the final model with 100% accuracy. Dark grey bars indicate the proportion of trees that voted the serovar was gastrointestinal, and light grey bars indicate the proportion of trees that voted the serovar was invasive. Out-of-bag votes are only those votes cast by trees that were not trained on a given sample.

Overall, the distribution of votes indicates that it is particularly difficult to find patterns unique to invasive isolates that are present in similar proportions in *S*. Dublin. This may be due to the greater range of potential hosts that *S*. Dublin can infect. Invasive and host-adapted salmonellae show a high degree of overlap, and the signal that we are picking up here may be a combination of signal from genetic changes related to invasiveness and changes related to being host adapted. Our top indicators show strong evidence of being related to invasiveness, as they contain genes that have been found to be important for metabolism in the inflamed gut environment (Nuccio & Baumler, 2014b; Rivera-Chávez & Bäumler, 2015). A subset of the genes picked up in the first iteration of the model could have been indicators of host restriction rather than invasiveness, thus confounding our results.

The OOB error rate uses trees not trained on a given serovar to predict whether it is invasive or gastrointestinal. Since *S*. Dublin and *S*. Enteritidis are so closely related, any tree with *S*. Dublin left out of its training data is likely to have *S*. Enteritidis left in and vice versa, so they will likely most closely mimic the score distribution of the other, leading to mis-classification. *S*. Gallinarum shows the most extensive gene degradation of all serovars, so is probably more exempt from this effect than the other two.

OOB error estimates give us an indication of how the model would perform identifying invasiveness in a serovar not used in the training data. In other words, it gives us an indication of how much of the signal captured is from general indicators of invasiveness rather than lineage specific markers. The voting proportions provided by the whole forest (Supplementary Figure 1) give us a better indication of how the model would perform if given another instance of a serovar included in this study. If information on genetic background is included, prediction accuracy is 100% even using the first model with no feature selection.

Cbi, ttr and pdu operons are represented as strong predictors

Among the top predictors were several sets of genes belonging to the same operon. Examples included the ttr, cbi and pdu operons. These operons in particular have been identified as key degraded pathways in invasive isolates (Thomson et al., 2008), and indicate the agreement of this method with other studies endeavouring to find links between loss of gene function and pathovar.

Most predictive genes show greater deviation from modelled sequence constraints in invasive strains

Of the top predictors in our study (N = 197), 154 showed significantly greater deviation from modelled sequence constraints in invasive strains compared to gastrointestinal strains (Mann-Whitney U test, adjusted *P*-value < 0.05), compared to 9 genes that showed greater deviation in gastrointestinal strains. Of the genes that were better conserved in invasive isolates, one was an aldo-keto reductase that was disrupted in all but one of our gastrointestinal strains and intact in all invasive strains, another was a chaperone protein, which we might expect to be more important in bacteria undergoing genomic degradation, particularly caused by increased drift due to population bottlenecking (Holt et al., 2009). Another was an aldose-1-epimerase that has been deleted from 4/7 of the gastrointestinal strains in our study.

Sequence changes in key indicator genes involve different mutations in each strain, contributing to similar functional outcomes

When looking at the individual genes that showed informative sequence differences between invasive and gastrointestinal isolates, we found that many of these changes had occurred independently, and had occurred at different sites in the protein. Figure 3 illustrates mutation accumulation in one of the top candidate genes, *mrcB*, a penicillin-binding protein required for bile tolerance (Langridge et al., 2009). Not only does the gene show more mutations in invasive strains compared to gastrointestinal strains, the mutations have occurred independently in different residues of the protein. Bile is thought to be an environmental signal recognised by *Salmonella* which causes suppression of invasiveness (Prouty & Gunn, 2000), possibly indicating the correct timing for initiation of invasive infection. While there are more mutations in invasive than in gastrointestinal strains, the mutations incur low scoring penalties according to the delta-bitscore metric used in this study to assess the functional impact of amino acid substitutions (Wheeler et al., 2016). This suggests that sequence changes could result in a change in protein function, rather than a loss. Consistent with this hypothesis, *mcrB* appears to be important for the survival of S. Typhi during a typical infection cycle (Langridge et al., 2009).

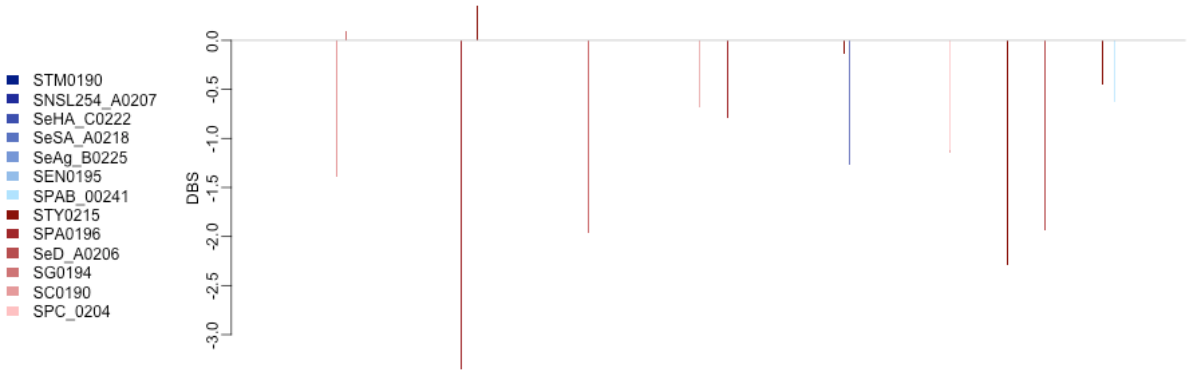


Figure 3 | Deleterious mutations in *mrcB*, one of the top three predictors
Mutations in different strains are colour-coded, with bars in red indicating a mutation in an invasive strain and bars in blue indicating a mutation in a gastrointestinal strain. An indication of the effect of the mutation on protein function is shown on the y-axis, with high negative values indicating higher chance of a mutation being deleterious to protein function. The x-axis represents the length of the protein.

Discussion

In this study we have employed a high-throughput, automated method for identifying patterns of mutation in protein coding genes associated with invasive pathovars of *Salmonella enterica*. Using this approach, we have identified genetic markers that have been identified by other studies as relevant to the transition between gastrointestinal and invasive infection mechanisms in *Salmonella* (Holt et al., 2009; Langridge et al., 2015; Nuccio & Baumler, 2014).

There are likely to be many genes involved in adaptation to invasiveness, but our final set of genes numbered just 197. This number was chosen because it provided perfect out-of-bag classification accuracy based on our training data, indicating that these genes capture trends related to invasiveness that were not unique to individual invasive strains in our study. Variable importance (VI) should rank genes by the level of parallelism seen in loss of function mutations, with a larger number of mutations seen in one group lending additional value to the gene as a predictor. This should help us to distinguish between genes that are under selection for loss or change of function, and genes that stochastically lose their function due to relaxed selection on the gene or increased drift in the population. Correlated predictors that are also correlated with outcome tend to be given higher VI, which is useful for our type of analysis, as it can help us identify pathways that are commonly affected and show parallel changes (Nicodemus et al., 2010). This may be a contributing factor to the appearance of multiple genes from the same operons in the top results.

We observed that prediction accuracy in earlier models was hampered by the difficulty in correctly identifying the phenotype of *S.* Enteritidis and *S*. Dublin. The random forest works by collecting cases of genes that tend to have diverged more in one group compared to the other, which is this case predominantly captures genes that are more likely to be degraded in an invasive isolate. Once this ensemble of trees containing strong discriminatory genes has been built, each serovar is then run through all of the trees that were not trained on them to produce a prediction of pathovar. If we used a larger dataset, OOB accuracy would likely have reached 100% earlier, as each OOB sample would be likely to contain one example from each invasive and each gastrointestinal pathovar, allowing the model to capture lineage specific mutational patterns indicative of each pathovar. This was not the outcome we wanted, however, so by using a much smaller training dataset we were able to identify patterns in a small collection of invasive pathovars which replicated in other invasive pathovars that had evolved independently. It is unusual to use random forests on such a

small dataset, but it gave us an opportunity to search for lineage independent signatures of invasiveness without employing any transformations of the data or corrections for population structure.

Clearly, moving forward it will be desirable to include more strains to train random forest models, and progress has been made in re-working the training mechanism of random forests to add in a random effect of population structure in the building of each node in a random forest to allow the inclusion of more strains without confounding effects of population structure (Stephan et al., 2015). Studies with a large number of strains will have more power to detect weaker associations, but the approach will turn into more of a “black box”, not because of lack of transparency, but because of the large amount of information to interpret. Keeping the training set small allowed us to explore the results more, and to look at the differences between the more distantly related strains, and the cluster of more closely related strains (*S*. Gallinarum, Enteritidis, Dublin). This preliminary study indicates that lineage effects will factor in strongly to the patterns detected that are associated with phenotype, and that correction for these is critical to identifying true genetic correlates with a phenotype.

References

Betancor, L., Yim, L., Martínez, A., Fookes, M., Sasias, S., Schelotto, F., … Chabalgoity, J. A. (2012). Genomic Comparison of the Closely Related Salmonella enterica Serovars Enteritidis and Dublin. *The Open Microbiology Journal*, *6*, 5–13.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees*. Chapman and Hall/CRC.

Holt, K. E., Thomson, N. R., Wain, J., Langridge, G. C., Hasan, R., Bhutta, Z. A., … Parkhill, J. (2009). Pseudogene accumulation in the evolutionary histories of Salmonella enterica serovars Paratyphi A and Typhi. *BMC Genomics*, *10*, 36.

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., … Bork, P. (2016). eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research*, *44*(D1), D286–93.

Klemm, E. J., Gkrania-Klotsas, E., Hadfield, J., Forbester, J. L., Harris, S. R., Hale, C., … Kingsley, R. A. (2016). Emergence of host-adapted Salmonella Enteritidis through rapid evolution in an immunocompromised host. *Nature Microbiology*, *1*, 15023.

Langridge, G. C., Fookes, M., Connor, T. R., Feltwell, T., Feasey, N., Parsons, B. N., … Thomson, N. R. (2015). Patterns of genome evolution that have accompanied host adaptation in Salmonella. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(3), 863–868.

Langridge, G. C., Phan, M.-D., Turner, D. J., Perkins, T. T., Parts, L., Haase, J., … Turner, A. K. (2009). Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. *Genome Research*, *19*(12), 2308–2316.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22.

Majowicz, S. E., Musto, J., Scallan, E., Angulo, F. J., Kirk, M., O’Brien, S. J., … International Collaboration on Enteric Disease “Burden of Illness” Studies. (2010). The global burden of nontyphoidal Salmonella gastroenteritis. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, *50*(6), 882–889.

Nicodemus, K. K., Malley, J. D., Strobl, C., & Ziegler, A. (2010). The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics*, *11*, 110.

Nuccio, S.-P., & Baumler, A. J. (2014). Comparative Analysis of Salmonella Genomes Identifies a Metabolic Network for Escalating Growth in the Inflamed Gut. *mBio*, *5*(2), e00929–14–e00929–14.

Pappu, V., & Pardalos, P. M. (2014). High-Dimensional Data Classification. In F. Aleskerov, B. Goldengorin, & P. M. Pardalos (Eds.), *Clusters, Orders, and Trees: Methods and Applications* (pp. 119–150). Springer New York.

Prouty, A. M., & Gunn, J. S. (2000). Salmonella enterica serovar typhimurium invasion is repressed in the presence of bile. *Infection and Immunity*, *68*(12), 6763–6769.

Rabsch, W., Andrews, H. L., Kingsley, R. A., Prager, R., Tschäpe, H., Adams, L. G., & Bäumler, A. J. (2002). Salmonella enterica serotype Typhimurium and its host-adapted variants. *Infection and Immunity*, *70*(5), 2249–2255.

Rivera-Chávez, F., & Bäumler, A. J. (2015). The Pyromaniac Inside You: Salmonella Metabolism in the Host Gut. *Annual Review of Microbiology*, *69*, 31–48.

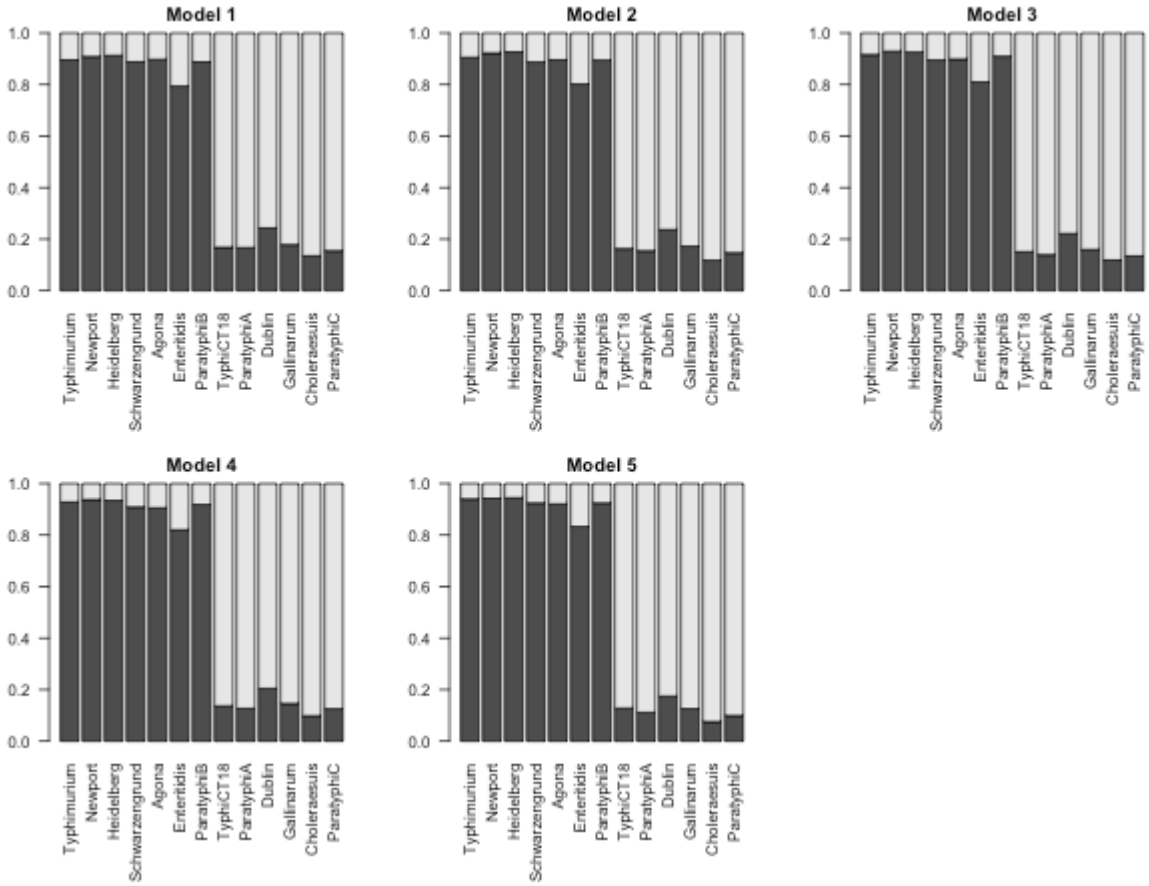
Stephan, J., Stegle, O., & Beyer, A. (2015). A random forest approach to capture genetic effects in the presence of population structure. *Nature Communications*, *6*, 7432.

Thomson, N. R., Clayton, D. J., Windhorst, D., Vernikos, G., Davidson, S., Churcher, C., … Parkhill, J. (2008). Comparative genome analysis of Salmonella Enteritidis PT4 and Salmonella Gallinarum 287/91 provides insights into evolutionary and host adaptation pathways. *Genome Research*, *18*(10), 1624–1637.

Touw, W. G., Bayjanov, J. R., Overmars, L., Backus, L., Boekhorst, J., Wels, M., & van Hijum, S. A. F. T. (2013). Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? *Briefings in Bioinformatics*, *14*(3), 315–326.

Wheeler, N. E., Barquist, L., Kingsley, R. A., & Gardner, P. P. (2016). A profile-based method for identifying functional divergence of orthologous genes in bacterial genomes. *Bioinformatics* , *32*(23), 3566–3574.

Supplementary Material



Supplementary Figure 1 | Votes of pathovar classification using the full random forest model

Supplementary Table 1 | Accession numbers of *Salmonella* enterica strains used in this study

Accession	Strain
AE006468	Typhimurium LT2
CP001113	Newport SL254
CP001120	Heidelberg SL476
CP001127	Schwarzengrund CVM19633
CP001138	Agona SL483
AM933172	Enteritidis P125109
CP000886	Paratyphi B SPB7
AL513382	Typhi CT18
CP000026	Paratyphi A ATCC 9150
CP001144	Dublin CT_02021853
AM933173	Gallinarum 287/91
AE017220	Choleraesuis SC-B67
CP000857	Paratyphi C RKS4594

Supplementary Table 2 | Top predictor genes

Gene name	Reference locus tag	Gene product
-	STM0018	Putative exochitinase
-	STM0019	Putative hydroxymethyltransferase
bcfC	STM0023	Fimbrial usher protein
-	STM0042	Putative sodium galactoside symporter
citF2	STM0061	Putative citrate lyase alpha chain
carB	STM0067	Carbamoyl-phosphate synthase large chain [EC=6.3.5.5]
caiC	STM0071	Probable crotonobetaine/carnitine-CoA ligase [EC=6.2.1.-]
araA	STM0102	L-arabinose isomerase [EC=5.3.1.4]
-	SeAg_B0155	Aldo-keto reductase YakC (NADP+) [EC=1.1.1.-]
hofC	STM0142	Putative component in type IV pilin biogenesis
mrcB	STM0190	Transpeptidase of penicillin-binding protein 1b
-	STM0257	Putative drug efflux protein (Perhaps for chloramphenicol)
leuD2	STM0330	3-isopropylmalate dehydratase small subunit 2 [EC=4.2.1.33]
-	STM0343	Putative diguanylate cyclase/phosphodiesterase domain 1
foxA	STM0364	Ferrioxamine B receptor
prpR	STM0367	Propionate catabolism operon regulatory protein
-	STM0382	Putative permease
phnX	STM0432	Phosphonoacetaldehyde hydrolase [EC=3.11.1.1]
yajL	STM0433	Chaperone protein YajL
cypD	STM0452	Peptidyl prolyl isomerase [EC=5.2.1.8]
mdlB	STM0461	Putative ABC superfamily (Atp&membrane) transporter
acrB	STM0475	RND family, acridine efflux pump
copA	STM0498	Copper-exporting P-type ATPase A [EC=3.6.3.n1]
ybbP	STM0508	Putative inner membrane protein
fimD	STM0546	Fimbrial usher protein
fimH	STM0547	Fimbrial tip adhesin protein
ybdZ	STM0587	Putative cytoplasmic protein
entE	STM0596	2,3-dihydroxybenzoate-AMP ligase [EC=6.3.2.-]
entA	STM0598	2,3-dihydro-2,3-dihydroxybenzoate dehydrogenase [EC=1.3.1.28]
ybdR	STM0615	Putative dehydrogenase
rna	STM0617	RNase I [EC=3.1.27.6]
ybeM	STM0631	Putative hydrolase
-	STM0652	Putative sigma-54 dependent transcriptional regulator
rihA	STM0661	Pyrimidine-specific ribonucleoside hydrolase RihA [EC=3.2.-.-]
tcuR	STM0692	Putative LysR family transcriptional regulator
pgm	STM0698	Phosphoglucomutase [EC=5.4.2.2]
kdpB	STM0705	Potassium-transporting ATPase B chain [EC=3.6.3.12]
ybfA	STM0708	Putative periplasmic protein
phrB	STM0709	Deoxyribodipyrimidine photo-lyase [EC=4.1.99.3]
pgl	STM0785	6-phosphogluconolactonase [EC=3.1.1.31]
ybhM	STM0808	Putative integral membrane protein
dinG	STM0821	Probable ATP-dependent helicase DinG [EC=3.6.4.12]
glnH	STM0830	Glutamine high-affinity transporter
ybiU	STM0841	Putative cytoplasmic protein
-	STY0946	Putative integrase
mukB	STM0994	Chromosome partition protein MukB
-	STM1060	Putative iron-sulfur protein
yccR	STM1072	Putative DNA transformation protein
helD	STM1075	DNA helicase IV [EC=3.6.1.-]
pipD	STM1094	Pathogenicity island encoded protein: SPI5
mdtG	STM1154	Multidrug resistance protein MdtG
htrB	STM1155	Lauroyl/myristoyl acyltransferase involved in lipid A biosynthesis [EC=2.3.1.-]
solA	STM1160	N-methyl-L-tryptophan oxidase [EC=1.5.3.-]
fhuE	STM1204	Outer membrane receptor for Fe(III)-coprogen
ycfR	STM1214	Putative outer membrane protein
nagK	STM1220	N-acetyl-D-glucosamine kinase [EC=2.7.1.59]
pepT	STM1227	Peptidase T [EC=3.4.11.4]
rluE	STM1237	Ribosomal large subunit pseudouridine synthase E [EC=5.4.99.20]
-	STM1252	Putative cytoplasmic protein
yeaG	STM1285	Putative Ser protein kinase
-	STM1330	Putative DNA/RNA non-specific endonuclease
pheT	STM1338	Phenylalanine--tRNA ligase beta subunit [EC=6.1.1.20]
pps	STM1349	Phosphoenolpyruvate synthase [EC=2.7.9.2]
ydiN	STM1360	Putative MFS family transport protein
ydiM	STM1361	Putative MFS family transport protein
ttrA	STM1383	Tetrathionate reductase subunit A [EC=1.8.-.-]
ttrC	STM1384	Tetrathionate reductase subunit C
ssaM	STM1413	Secretion system apparatus protein SsaM
rnfC	STM1457	Electron transport complex protein RnfC
clcB	STM1490	Voltage-gated ClC-type chloride channel ClcB
-	STM1498	Putative dimethyl sulphoxide reductase [EC=1.8.99.-]
-	STM1499	Putative dimethyl sulphoxide reductase, chain A1 [EC=1.8.99.-]
ynfD	STM1500	Putative outer membrane protein
dcp	STM1512	Peptidyl-dipeptidase dcp [EC=3.4.15.5]

ydeE	STM1516	Putative MFS family transport protein
pqaA	STM1544	PhoPQ-regulated protein
-	STM1545	Putative multidrug efflux protein
-	STM1546	Putative monooxygenase
-	STM1547	Putative marR-family transcriptional regulator
-	STM1549	Putative translation initiation inhibitor
-	STM1556	Putative Na+/H+ antiporter
ydcP	STM1604	Putative collagenase [EC=3.4.-.-]
-	STM1615	Putative nucleoside triphosphatase
-	STM1624	Putative cytoplasmic protein
trg	STM1626	Methyl-accepting chemotaxis protein III
-	STM1630	Putative inner membrane protein
oppA	STM1746.S	Periplasmic oligopeptide-binding protein
hyaA	STM1786	Hydrogenase-1 small subunit [EC=1.12.7.2]
nhaB	STM1806	Na(+)/H(+) antiporter NhaB
-	SG1259	Putative uncharacterized protein
-	SG1196	Putative phage encoded hydrolase
-	SG1195	Predicted phage protein
mrdA	STM1910	Penicillin-binding protein 2
flhB	STM1914	Flagellar biosynthetic protein FlhB
-	STM1940	Putative cell wall-associated hydrolase
cbiK	STM2025	Sirohydrochlorin cobaltochelatase [EC=4.99.1.3]
cbiD	STM2032	Putative cobalt-precorrin-6A synthase [deacetylating] [EC=2.1.1.-]
cbiA	STM2035	Cobyrinic acid A,C-diamide synthase
pocR	STM2036	Regulatory protein PocR
pduG	STM2043	Propanediol utilization diol dehydratase reactivation protein
pduS	STM2053	Propanediol utilization protein
pduW	STM2057	Probable propionate kinase [EC=2.7.2.15]
phsA	STM2065	Thiosulfate reductase [EC=1.-.-.-]
sopA	STM2066	E3 ubiquitin-protein ligase SopA [EC=6.3.2.-]
wcaJ	STM2103	Putative UDP-glucose lipid carrier transferase
mdtC	STM2128	Multidrug resistance protein MdtC
mglA	STM2189	Galactose/methyl galactoside import ATP-binding protein MglA [EC=3.6.3.17]
yejF	STM2219	Putative ATPase component of ABC-type transport system
-	STM2245	Putative outer membrane protein
napA	STM2259	Periplasmic nitrate reductase [EC=1.7.99.4]
sseL	STM2287	Deubiquitinase SseL [EC=3.4.22.-]
rhmA	STM2289	2-keto-3-deoxy-L-rhamnonate aldolase [EC=4.1.2.n3]
yfaZ	STM2294	Putative inner membrane protein
nuoN	STM2316.S	NADH-quinone oxidoreductase subunit N [EC=1.6.99.5]
-	STM2357	Putative amino acid transporter
-	STM2359	Putative amino acid transporter
yfcJ	STM2372	UPF0226 protein YfcJ
fadJ	STM2388	Fatty acid oxidation complex subunit alpha
pgtA	STM2396	Phosphoglycerate transport system transcriptional regulatory protein PgtA
-	STM2405	Putative thiamine pyrophosphate enzymes [EC=4.1.1.1]
nupC	STM2409	NUP family nucleoside transport protein
yfeA	STM2410	Putative diguanylate cyclase/phosphodiesterase domain 1 containing protein
eutR	STM2454	HTH-type transcriptional regulator EutR
eutH	STM2460	Ethanolamine utilization protein EutH
aegA	STM2479	Putative oxidoreductase
acrD	STM2481	RND family aminoglycoside/multidrug efflux pump
-	STM2503	Putative diguanylate cyclase
-	STM2529	Putative anaerobic dimethylsulfoxide reductase
-	STM2532	Putative inner membrane lipoprotein
purL	STM2565	Phosphoribosylformylglycinamidine synthase [EC=6.3.5.3]
mltF	STM2567	Membrane-bound lytic murein transglycosylase F [EC=4.2.2.n1]
nadB	STM2641	L-aspartate oxidase [EC=1.4.3.16]
nrdF	STM2808	Ribonucleoside-diphosphate reductase 2 subunit beta [EC=1.17.4.1]
emrB	STM2815	Putative MFS superfamily multidrug transport protein
norV	STM2840	Anaerobic nitric oxide reductase flavorubredoxin
hycC	STM2851	Hydrogenase 3, membrane subunit
ygcY	STM2961	Putative d-glucarate dehydratase [EC=4.2.1.40]
ygdH	STM2969	Putative nucleotide binding
yqeF	STM3019	Putative acetyl-CoA acetyltransferase [EC=2.3.1.9]
-	STM3021	Putative inner membrane protein
rcnA	STM3024	Nickel/cobalt efflux system RcnA
serA	STM3062	D-3-phosphoglycerate dehydrogenase [EC=1.1.1.95]
-	STM3073	Putative ABC-type cobalt transport system
-	STM3074	Putative ABC-type cobalt transport system
-	STM3081	Putative malate
-	STM3083	Putative mannitol dehydrogenase
yggM	STM3105	Putative periplasmic protein
-	STM3125	Putative cytoplasmic protein
mcpB	STM3152	Putative methyl-accepting chemotaxis protein
ygjR	STM3223	Putative dehydrogenase
yrbB	STM3309	Putative STAS domain protein

-	STM3355	Putative tartrate dehydratase alpha subunit [EC=4.2.1.32]
smf	STM3405	Putative uncharacterized protein smf
kefB	STM3457	Glutathione-regulated potassium-efflux system protein KefB
damX	STM3485	Membrane protein
livJ	SG3870	Leu/ile/val/thr-binding protein
yhjC	STM3607	Putative LysR family transcriptional regulator
bcsA	STM3619	Cellulose synthase catalytic subunit [UDP-forming] [EC=2.4.1.12]
lpfC	STM3638	Fimbrial usher protein
bisC	STM3644	Biotin sulfoxide reductase [EC=1.-.-.-]
malS	STM3664	Alpha-amylase [EC=3.2.1.1]
lyxK	STM3674	Cryptic L-xylulose kinase [EC=2.7.1.53]
sgbU	STM3676	Putative 3-hexulose-6-phosphate isomerase [EC=5.-.-.-]
rfaC	STM3712	Lipopolysaccharide heptosyltransferase 1 [EC=2.-.-.-]
rfaZ	STM3715	Lipopolysaccharide core biosynthesis protein RfaZ
slsA	STM3761	Putative inner membrane protein
mgtB	STM3763	Magnesium-transporting ATPase, P-type 1 [EC=3.6.3.2]
-	STM3773	Putative NtrC family transcriptional regulator
uhpB	STM3789	Sensor protein UhpB [EC=2.7.13.3]
yihQ	STM4019	Putative alpha-xylosidase
-	STM3834	Putative LysR family transcriptional regulator
yidZ	STM3848	HTH-type transcriptional regulator YidZ
-	STM3858	Putative phosphotransferase system fructose-specific component IIB [EC=2.7.1.69]
-	STM3859	Putative shikimate [EC=1.1.1.25]
gppA	STM3913	Guanosine-5'-triphosphate,3'-diphosphate pyrophosphatase [EC=3.6.1.40]
yihR	STM4020.S	Putative aldose-1-epimerase
-	STM4031	Putative cytoplasmic protein
fdoH	STM4036	Formate dehydrogenase-O, Fe-S subunit [EC=1.2.1.21]
priA	STM4095	Primosomal protein N'
pepE	STM4190	Peptidase E [EC=3.4.13.21]
tyrB	STM4248	Aromatic-amino-acid aminotransferase [EC=2.6.1.57]
dcuS	STM4304	Sensory histidine kinase in two-component retgulatory system with DcuR [EC=2.7.3.-]
dsbD	STM4323	Thiol:disulfide interchange protein DsbD [EC=1.8.1.8]
yjeH	STM4328	Putative transporter
yjeJ	STM4332	Putative inner membrane protein
blc	STM4339	Outer membrane lipoprotein
yjeF	STM4356	Putative sugar kinase
aidB	SEN4143	Probable acyl Co-A dehydrogenase
-	STM4413	Putative imidazolonepropionase and related amidohydrolases
-	STM4519	Putative NAD-dependent aldehyde dehydrogenase [EC=1.2.1.16]
-	STM4534	Putative NtrC family transcriptional regulator, ATPase domain protein
yjjI	STM4566	Putative cytoplasmic protein
deoB	STM4569	Phosphopentomutase [EC=5.4.2.7]
lplA	STM4576	Lipoate-protein ligase A [EC=2.7.7.63]
creB	STM4588	Response regulator in two-component regulatory system with CreC
creD	STM4590	Tolerance to colicin E2 protein

Chapter Six | Draft: Functional adaptation of the core genome in *Pseudomonas* plant pathogens

Preface

So far we have looked at applications of the delta-bitscore metric to within-species comparisons. We expect this to be a strength of the software, because sequence changes to protein coding genes that occur over a shorter time span are more likely to involve a small number of mutations. This will cause the sequence differences we are testing to more closely resemble the data we used for benchmarking our method, which contained only one substitution per sequence. There is still uncertainty as to how the delta-bitscore method will perform when given sequences that diverged a long time ago and have acquired a larger number of mutations. Over time, the 3D context surrounding any mutated residue may change, we may see compensatory mutations occur over evolutionary time, which could be modelled poorly by a hidden Markov model.

In an attempt to assess the viability of using this approach for between-species comparisons, we used the method to compare the genomes of *Pseudomonas* plant pathogens with those of harmless or beneficial rhizosphere-associated and environmental *Pseudomonas* isolates. The approach identified sequence divergence of potential functional importance in a number of proteins that had previously been implicated in pathogenicity in *Pseudomonas* species. Further experimental work is required to verify the importance of some of these genes in pathogenicity, and more scepticism should be applied to analysis of protein coding sequences at this level of evolutionary distances than to the analysis of closely-related isolates, however the availability of good plant infection models for *P. syringae* means that we are able to test these candidate genes in the future for an effect on plant infection outcomes.

Contributions

Honour McCann selected the strains for analysis and contributed to the introduction, I performed DBS analysis and wrote most of the manuscript. Paul Gardner contributed to the design of the analysis.

Functional adaptation of the core genome in *Pseudomonas* plant pathogens

Nicole E. Wheeler[1], Paul P. Gardner[1], Honour McCann[2]

1. School of Biological Sciences, University of Canterbury, Christchurch, New Zealand.
2. New Zealand Institute for Advanced Study and Allan Wilson Centre, Massey University, Auckland, New Zealand

Abstract

Pseudomonas spp. are found in a variety of niches, owing to their broad metabolic capabilities and ability to adapt to a wide range of environments. *Pseudomonas* invasion of plant hosts is of particular importance and interest. Previous studies have identified the Type 3 secretion system and secreted virulence factors as playing a major role in pathogen virulence, yet successful invasion and proliferation in this niche is likely accompanied by additional adaptations. We employ a novel approach to identify different patterns of functionally significant sequence variation in genes from plant pathogenic *Pseudomonas* spp. This profile hidden Markov model-based method identifies genes previously linked to pathogenicity as well as numerous un/poorly characterised genes. Our data reveals numerous candidate proteins which may be linked with adaptation to the plant environment, and reinforce the need to investigate the contribution of non-T3SS-related genes to pathogenicity.

Introduction

Despite the ready availability of genomic data for an increasing number of microbial pathogens, the identification of the genetic determinants of pathogenicity remains an elusive goal for non-model pathosystems. This is partly because the rate of sequencing exceeds the speed with which new analytical tools are sufficiently developed to provide researchers with compelling candidates for functional characterisation. Transposon mutagenesis screens are a popular discovery-based approach for the identification of genes required for growth and colonisation (Meziane et al., 2014). These will reveal candidates that are essential for virulence (secretion systems and genes involved in regulatory pathways, for example), but fail to provide meaningful insight into the genomic changes that accompany adaptation to the pathogenic lifestyle.

Pseudomonas are a ubiquitous Gram-negative proteobacteria that occupies a broad range of ecological niches. Pseudomonads are present in the soil (e.g. *P. stutzeri*), form biofilms in freshwater environments (e.g. *P. syringae*), reside on leaf surfaces (e.g. *P. fluorescens*) and are pathogenic on plants (e.g. *P. syringae*), basidiomycete mushrooms (e.g. *P. tolaasii*) and immunocompromised humans (e.g. *P. aeruginosa*) (Brodey, 1991; Driscoll et al., 2007; Hebbar et al., 1991; Lalucat et al., 2006; Morris et al., 2008). The apparent ubiquity of this clade has been attributed to their versatile metabolism and their ability to adapt to a range of environmental conditions and stressors (Silby et al., 2011).

Pseudomonas syringae serves as one of the most important model systems for studies in plant pathogen evolution and molecular plant-microbe interactions (O'Brien et al., 2011). *P. syringae* is a large species complex comprising nonpathogenic environmental isolates as well as around 60 different pathogen variants (pathovars) exhibiting highly specific interactions with distinct plant hosts and cultivars (Marcelletti & Scortichini, 2014). The host of isolation (from which a strain is assigned a pathovar name) is generally the host upon which that strain is most virulent, though host range tests across a broad sampling of plant species are not routinely performed. Significant efforts have focused on identifying the genetic basis of virulence and host specificity in *P. syringae*. Virulence proteins secreted by the Type 3 secretion system (T3SS) are considered the most important virulence determinants: knocking out the structural components of the T3SS eliminates the pathogen's ability to suppress and subvert host recognition responses (Hye-Sook Oh et al., 2010; Zumaquero et al., 2010).

Not all *Pseudomonas* spp. colonizing plant tissues are pathogenic however: nonpathogenic strains have been isolated from the rhizosphere as well as the surface and interior of aerial plant tissues (Hebbar et al., 1991; Preston, 2004; Wahyudi et al., 2011). It is not yet clear whether these simply persist until there is opportunity for infection, are unable to establish infections, or have been effectively suppressed by host defenses. The colonisation and invasion of plant hosts is a complex interaction, and is likely accompanied by multiple adaptations in addition to the acquisition of a T3SS and assorted effectors. Although these adaptations underpin a major transition from free-living generalism to specialisation on plants, they may not be detected by screening knockout libraries if their function is required during a stage of the pathogen life cycle not assayed in conditional essentiality experiments.

We hypothesise that the selective pressures particular to the plant niche result in differences in patterns of functionally significant genetic variation. More specifically, we propose that pathogenic and free-living non-pathogenic microbes display different patterns of accumulated mutations altering protein function. Further, we propose that these patterns can be exploited to identify genes involved in niche adaptation or virulence, and predict whether a bacterium is more likely to be a free-living generalist or a plant pathogen. Insights into these differences will elucidate the genetic and functional requirements for pathogenicity in *Pseudomonas* and provide us with novel ways of detecting *Pseudomonas* plant pathogens.

We employed a novel technique developed by Wheeler et al (2016) based upon profile HMMs to identify functionally significant genetic variation in orthologs shared by 9 pathogenic and 18 non-pathogenic *Pseudomonas* spp.. Profile HMMs are a powerful homology-search tool (Eddy, 2011), that we have adapted for evaluating the significance of genetic variation and distinguishing functionally important sequence variation from functionally neutral or near-neutral changes. The application of this novel approach allows for the identification of non-synonymous changes in protein coding genes that defy the patterns of sequence variation usually observed in homologous sequences. Thus, we are able to identify changes that occur in these lineages but have been underrepresented in homologs of our proteins of interest, indicating they are likely to be disruptive of protein structure or function.

In this analysis we have identified different sequence constraints associated with pathogenicity in a range of genes that have been associated to some degree with pathogenicity in the literature, but also genes of unknown or poorly characterised function that may be associated with a pathogenic lifestyle, suggesting a need for further study of these candidates in order to understand whether they play important roles in pathogenicity.

Methods

In this analysis, we have collected genome sequences from a range of *Pseudomonas* species from well characterised ecological niches that can be classified as plant pathogens, rhizosphere associated, or environmental with no known interaction with plants. We examined the suite of protein coding genes shared by most members of pathogenic and nonpathogenic groups in order to identify genes that showed differences in the degree to which they adhered to sequence constraints typically observed in these proteins.

Genome sequences

Genome sequences for 27 *Pseudomonas* genomes were downloaded from the NCBI website (ftp://ftp.ncbi.nih.gov/genomes/genbank/bacteria/). See Supplementary Table 1 for full list of strains and accession numbers. Bacteria were divided into three groups: pathogenic, rhizosphere associated and environmental, based on information in the literature.

Ortholog identification and profile-based analysis

Orthologs were identified using profile hidden Markov models (HMMs) built from gamma proteobacterial genomes retrieved from the eggNOG profile HMM collection (Huerta-Cepas et al., 2016). Each protein sequence was searched against this HMM database using hmmsearch from the HMMER3.0 package (http://hmmer.org). Hits to eggNOG models were filtered using an E-value cutoff of 0.0001, then the top scoring protein corresponding to each model was used for analysis. Genes were filtered based on their presence in isolates from each group (present in at least 5 pathogens and 10 nonpathogens). A significant difference in distribution between groups, as measured by a Kolmogorov-Smirnov test, as well as a difference in median bitscore (here deemed delta-bitscore, or DBS) between groups that fell in the most extreme 5% ($|DBS| > 37.85$) were required for initial screening for functional divergence. The genes that scores in the most extreme 1% ($|DBS| > 111.95$) were also excluded from analysis, as these were likely to contain a high proportion of mis-called orthologs. This gave us a shortlist of 87 genes. After further manual filtering for mis-called orthologs, 51 genes remained.

In order to assess the background level of functional divergence that occurs between groups of organisms, we repeated the analysis with two phylogenetic subgroups identifiable in the species included in our study (Supplementary Figure 1), each containing isolates of mixed phenotype. For this analysis, orthologous groups were only tested if the gene was present in at least ten members from each group, and the top 1% of delta-bitscores were removed to exclude the most likely mis-called orthologs.

Results

In this study we have identified a subset of genes shared by pathogenic and nonpathogenic pseudomonads, which show differences in adherence to patterns of sequence conservation observed in homologs from a range of related species.

A subset of genes show distinct score distributions for pathogens and nonpathogens

In examining score distributions for the orthologous groups defined by eggNOG family membership, we found that while many gene families showed high variance in the degree to which individual sequences fit sequence constraints captured by the eggNOG protein family models (as measured by bitscore values for each match), only a subset of genes (N = 87) showed differences in score associated with a pathogenic lifestyle (Figure 1). This indicates that a substantial amount of functionally important sequence space has been explored over time, however only a subset of genes show greater adherence to modelled sequence constraints in pseudomonads leading a pathogenic lifestyle compared to a non-pathogenic lifestyle or vice-versa. 51 genes were identified that fit our criteria for showing different sequence constraints in the two groups after manual examination of sequences to exclude mis-called orthologs. Among these genes, we were able to identify functional categories that may be associated with adaptation to a pathogenic niche.

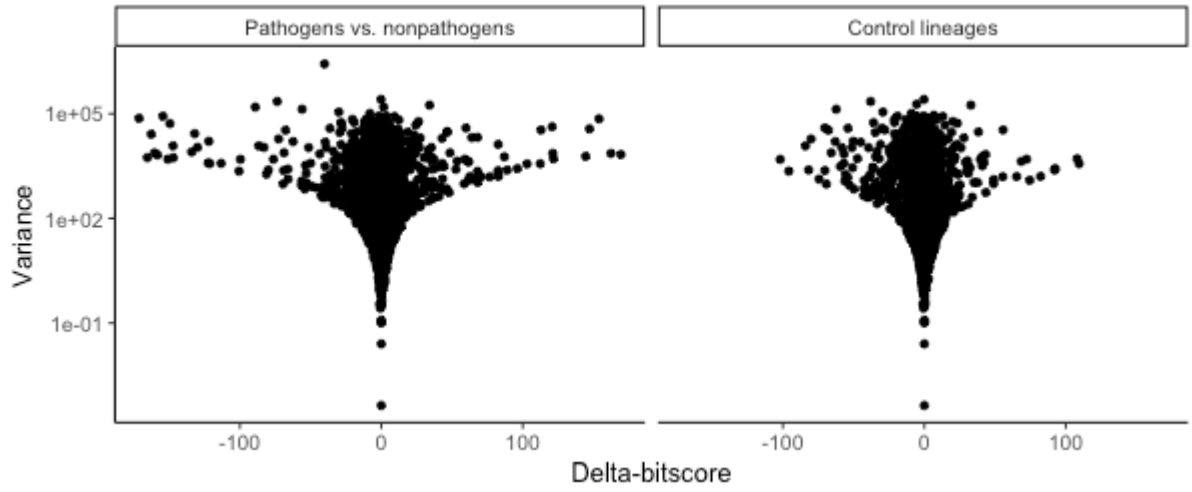


Figure 1 | While agreement of individual sequences within an orthologous group to modelled sequence constraints for that protein family can vary widely, only a small subset of groups show directional variation associated with lifestyle

Shown are delta-bitscore values vs variance for orthologous groups after filtering out the top 1% of DBS values. The left-hand panel shows a comparison of pathogenic and nonpathogenic pseudomonads, the right-hand panel shows a comparison of another monophyletic group of mixed phenotype to other pseudomonads. We see that functional divergence specific to pathogens is greater than that seen for a lineage of mixed phenotype.

When we examined another monophyletic group within our study set, to test whether comparing one lineage to a broader sampling of pseudomonads will always produce such a result, we found that the spread of delta-bitscores was far smaller (Figure 1). This indicates

that the sequence divergence we see in proteins from the predominantly monophyletic pathogenic groups in our study was greater than the effect we might expect as an artefact of comparing a group of closely related species with outgroups.

Functions of top candidate genes can be linked to requirements for a pathogenic lifestyle

We observed a mixture of gene families that had higher and lower scores in the pathogens compared to the nonpathogens, indicating that there was no strong signal of gene degradation in pathogenic bacteria that could be captured on this timescale using this metric. This finding was somewhat surprising given the importance of gene degradation to pathogenicity reported in the literature (Maurelli, 2007; Merhej et al., 2013). Many of the genes we identified are of unknown or poorly characterised function. Of those with known functions, several themes emerged that may relate to differences in selection pressures associated with a pathogenic lifestyle (see Supplementary Table 2 for full list of candidate genes).

Type III secretion system

The genes encoding CbrA, FliH and a lytic murein transglycosylase all had higher scores in pathogenic pseudomonads. CbrA is known to be required for symbiosis by nitrogen-fixing microbes, regulating motility and chemotaxis, metabolism and cell envelope function (Gibson et al. 2007; Gibson et al., 2006), but is also involved in metabolic regulation of the T3SS (Yeung et al., 2011). Flagellar assembly is thought to antagonise T3SS activity in *Pseudomonas* (Soscia et al., 2007), but is negatively regulated by FliH. The lytic murein transglycosylase is thought to be a contributor to the translocation of effectors into plant cells (Oh et al., 2007). These findings all support previous research that T3SS genes are important to the pathogenicity of some pseudomonads, even when these genes can be found in both pathogens and nonpathogens.

Siderophore biosynthesis, iron uptake

We observed sequence changes associated with pathogenicity in proteins involved in metal acquisition and transport. Specifically, the proteins EfeO, HemY, and PvdN had higher scores in pathogens than in non-pathogens, while FecR showed the reverse pattern. HemY, PvdN and FecR are involved in siderophore synthesis, and both PvdN and FecR have also been implicated in virulence (Lamont et al., 2002; Taguchi et al., 2009). EfeO is believed to be involved in iron transport (Rajasekaran et al., 2010). In pathogenic interactions between

plant and bacteria, siderophores are required for bacterial survival, and often also for full expression of virulence (Aznar & Dellagi, 2015).

Reactive oxygen species metabolism

The production of reactive oxygen species is a well-known early response to pathogen recognition by plants. This response is thought to be involved in both killing invading pathogens and signalling subsequent responses in the plant (for example, changing gene expression patterns) (Torres, 2006). Nitric oxide is a reactive oxygen species (ROS) involved in plant immune responses, but nitric oxide metabolism is also a key part of bacterial survival in a plant host (Arasimowicz-Jelonek et al., 2013).

Hmp, a mannitol reductase gene and a glutathione-S-transferase gene all had higher scores in pathogens compared to nonpathogens. Hmp is an NO-inducible flavohemoprotein which has been found to confer a protective effect against killing by macrophages in *E. coli* by acting as a nitric oxide detoxifying protein (Stevanin et al., 2007). Mannitol is a potent ROS quencher which is synthesised by many pathogens to protect against the plant immune system (Jennings et al., 1998). Mannitol is also produced by *P. putida* under osmotic stress (Kets et al., 1996). The glutathione-S-transferase gene was of poorly characterised function, however these genes are known to have diverse roles in response to oxidative stress (Veal et al., 2002).

Some bacteria, notably bacterial plant pathogens, produce NO in response to cues indicating contact with a plant host (Johnson et al., 2008). Bacterial NO production is thought to serve a different role to that of eukaryotic NO production, functioning more in the nitration of different substrates and protection from oxidative stress (Sudhamsu & Crane, 2009). A flavodoxin nitric oxide synthase showed higher scores in nonpathogens compared to pathogens. Its function has not been well characterised, so the precise role it plays in host-microbe interaction is not known.

Cyclic di-GMP signalling

Cyclic di-GMP (c-di-GMP) is a bacterial second messenger that has been shown to regulate a variety of virulence-relevant traits, such as motility, virulence factor production, biofilm formation, as well as more fundamental processes such as the cell cycle and differentiation (Römling et al., 2013). Most signalling pathways that use c-di-GMP are involved in mediating interaction of the bacteria with surfaces, or other bacterial or eukaryotic cells (Römling et al.,

2013). As such, proteins controlling c-di-GMP production are not a surprise as top candidates for determining pathogenicity.

RbdA, a diguanylate cyclase phosphodiesterase which has been implicated in biofilm formation and regulation of motility (An et al., 2010; Newell et al., 2011) had higher scores in nonpathogens than in pathogens. Knockout of *rbdA* has been shown to reduce biofilm dispersion in *P. aeruginosa*, downregulating biofilm formation and upregulating factors associated with biofilm dispersion (An et al., 2010). In contrast, a diguanylate cyclase had higher scores in pathogens than in non pathogens.

c-di-GMP can be involved in the transition from a highly virulent state to a less virulent state conducive to chronic infections, so loss of gene function could result in a more virulent pathogen (Römling et al., 2013). c-di-GMP is also recognised by mammalian immune systems as a uniquely bacterial molecule (Römling et al., 2013), and therefore unnecessary production could be selected against in pathogens in order to avoid detection. *Y. pestis* has been found to carry a number of psedogenized enzymes for synthesizing c-di-GMP, and evidence indicates that c-di-GMP production by these genes was deleterious for virulence (Bobrov et al., 2011). Free-living bacteria tend to carry more enzymes for the production of c-di-GMP in order to cope with more varying environments (Römling et al., 2013), therefore we may expect some loss of these genes with a more stable environment.

Biofilm formation

In addition to c-di-GMP being a key regulator of biofilm formation, several other genes that contribute to biofilm formation also showed key differences in their adherence to modeleld sequence constraints. WcaJ, an extracellular polysaccharide biosynthesis protein required for biofilm formation (Yu et al., 2015), had higher scores in pathogens. CheA and AdnA are chemotaxis proteins that regulate biofilm formation (Mastropaolo et al., 2012; Robleto et al., 2003; Tremaroli et al., 2011). Both genes showed higher scores in nonpathogens compared to pathogens.

Resistance to drugs and toxins

Scores for the multidrug efflux protein MexD and a microcin C transporter were higher in pathogens than in non-pathogens. A fusaric acid resistance protein showed higher scores in nonpathogens compared to pathogens. Fusaric acid is produced by fungi of the family Fusarium, and is toxic to both plants and rhizobacteria. Many *Pseudomonas* can prevent

fusaric acid induced plant wilting by secreting siderophores such as pyoverdine (Ruiz et al., 2015). It may be that this function is important for mutualists, but pathogens avoid exposure due to differences in their localization in the host, so do not require production of a resistance protein. Interestingly, this protein was found to be present and scored highly in five environmental species as well as in the rhizosphere associated isolates. Scores for OmpF were higher for pathogens, even though this large porin can increase susceptibility to antibiotics (Kishii & Takei, 2009).

Nutrient acquisition

We found differences in genes regulating the transport and metabolism of different carbon sources. CdaR and a glycerol transporter had higher scores in nonpathogenic pseudomonads. CdaR is a diacid biosensor regulating enzymes that metabolise glucarate, galactarate and glycerate. A putative transporter of pectin lyase had higher scores in the pathogens in our study. Amino acid transporters also showed signs of functional change between pathogens and nonpathogens, with an L-glutamate/L-aspartate binding protein having higher scores in pathogens, and a D-methionine transporter having higher scores in nonpathogens.

Discussion

Our analysis presents a unique contribution to the study of pathogenicity in *Pseudomonas*. Previous analyses have focused on shorter timescales and examined specific pathogens, while we have taken a broader view of a group of pathogens and what distinguishes them from their relatives. Medina and Sachs (2010) point out that many factors that are useful for virulence (e.g. T3SS) are also used by nonvirulent and mutualist symbionts. Thus, we wanted to look at both environmental and beneficial pseudomonads in comparison to plant pathogens to identify functions that were uniquely maintained or degraded in the pathogens. Merhej et al (2013) observed that genome reduction is a common and important factor in pathogen evolution, so we endeavoured to not only find genes whose function appeared to be important in pathogens but not nonpathogens, but also to find genes that were degraded in pathogens specifically.

In our analysis a number of genes were determined to be under different sequence constraints between the pathogenic and nonpathogenic groups. Interestingly, we did not find an over-representation of degraded genes in the pathogenic group compared to the nonpathogens, as we had anticipated. Of those genes that were of known or predicted

function, several themes emerged which are consistent with what we might expect for functional requirements of *Pseudomonas* plant pathogens. We also identified a number of genes annotated as hypothetical proteins which, to the best of our knowledge have not yet been experimentally characterised in *Pseudomonas*. Many other genes were annotated according to strong Pfam domain architecture hits, however this provides limited value in ascertaining the role a protein may play in pathogenesis, particularly when a domain, or domain architecture is known to be found in proteins with diverse roles. This does, however present clues as to potential observable phenotypes that could be tested for some of these genes of unknown function. Functional annotation of proteins of unknown function has been identified as a key step toward identification of more druggable target proteins in pathogenic organisms (Ravooru et al., 2014). Due to the vast number of uncharacterised proteins present in many genomes, functional characterisation of such proteins can be a daunting task, but prioritisation of candidates for further study can help make the task more manageable (Galperin & Koonin, 2004). These findings present a link between a subset of these proteins and a key phenotype of interest, pathogenicity, suggesting that these proteins should be given higher prioritisation for functional characterisation, and suggesting that pathogenicity assays may be an effective way of testing for a change of phenotype resulting from knockout of these genes.

There are a few caveats in the interpretation of these results that the reader should be aware of. Because the models we have used are built using sequences collected from a range of gamma proteobacteria, the functions of these proteins may be broad in some cases. As such, deviation from modelled sequence constraints could indicate degradation of ancestral function, or it could indicate a shift in function from that of the shared ancestral protein. Because of these contrasting possibilities, while these results can indicate a difference in sequence constraints associated with the two lifestyles, we can not draw strong conclusions about whether these changes relate to loss or gain of function without experimental characterisation of the proteins.

A potential limitation to the interpretation of these changes as adaptive in a pathogenic lifestyle is the lineage-specific pathogenic phenotype. Most of the plant pathogens in our analysis come from a single monophyletic lineage. As a result, many of the trends we find will be unique to this lineage, and difficult to link directly to pathogenicity rather than other events that may have occurred along that branch of the tree. A promising finding was that while variance in scores could be very high for some genes, genes with high DBS were less

common. This indicated that sequence variation deemed to be functionally significant was only over- or under-represented in pathogens for a small subset of total genes that we investigated. In order to assess the background level of functional divergence that occurs between groups of organisms, we repeated the analysis with two subgroups identifiable in the tree of all of the species. One subgroup contained isolates of a range of phenotypes, with a skew toward rhizosphere-associated isolates, the other contained approximately even numbers of pathogens and environmental isolates. In this comparison, we saw that while the variance in scores was in the same range, the range of delta-bitscores was far smaller, indicating lineage-specific deviation in bitscores was more subtle when looking at a lineage with members occupying a combination of different niches. This lends support to our belief that the difference in adherence to modelled sequence constraints we see in our list of candidate genes is related to pathogenic phenotype rather than lineage-specific divergence in protein sequence unrelated to niche adaptation.

The candidate genes we have identified show signs that they may be under different sequence constraints in pathogens and nonpathogens. The screen performed in this study will be further refined by testing for functional differences between *P. syringae* isolates which occupy the same kiwifruit host either as hemibiotrophic pathogens or as symptomless endophytes. Genes displaying significant delta-bitscores will be more specifically predictive of pathogenicity on kiwifruit. Strains of both nonpathogenic and pathogenic *P. syringae* with knockouts in select candidate genes will then be tested to assess whether loss has fitness consequences for growth in vitro or in planta. If this screen provides promising results, allele exchange can be used to assess whether a pathogenic allele confers high growth in planta to a non-pathogenic strain, and vice versa.

Our analysis to date presents a collection of genes that appear to be under different sequence constraints in pathogenic pseudomonads compared to bacteria that either form mutualistic interactions with plants or are not known to interact with plants. We have identified proteins that appear to be undergoing divergence from ancestral sequence constraints in *Pseudomonas* plant pathogens, as well as genes that appear to be more carefully maintained in pathogens, suggesting importance for pathogenicity. Our analysis suggests a number of poorly characterised proteins may be playing important roles in pathogenicity, indicating a need for greater study of their function. These results will be important for the understanding of host-pathogen interaction in *Pseudomonas*, as well as for the search for new druggable targets to improve disease control strategies.

References

An, S., Wu, J. 'en, & Zhang, L.-H. (2010). Modulation of Pseudomonas aeruginosa biofilm dispersal by a cyclic-Di-GMP phosphodiesterase with a putative hypoxia-sensing domain. *Applied and Environmental Microbiology*, *76*(24), 8160–8173.

Arasimowicz-Jelonek, M., Magdalena, A.-J., & Jolanta, F.-W. (2013). Nitric oxide: an effective weapon of the plant or the pathogen? *Molecular Plant Pathology*, *15*(4), 406–416.

Aznar, A., & Dellagi, A. (2015). New insights into the role of siderophores as triggers of plant immunity: what can we learn from animals? *Journal of Experimental Botany*, *66*(11), 3001–3010.

Bobrov, A. G., Kirillina, O., Ryjenkov, D. A., Waters, C. M., Price, P. A., Fetherston, J. D., … Perry, R. D. (2011). Systematic analysis of cyclic di-GMP signalling enzymes and their role in biofilm formation and virulence in Yersinia pestis. *Molecular Microbiology*, *79*(2), 533–551.

Brodey, C. L. (1991). Research Notes Bacterial Blotch Disease of the Cultivated Mushroom Is Caused by an Ion Channel Forming Lipodepsipeptide Toxin. *Molecular Plant-Microbe Interactions: MPMI*, *4*(4), 407.

Driscoll, J. A., Brody, S. L., & Kollef, M. H. (2007). The epidemiology, pathogenesis and treatment of Pseudomonas aeruginosa infections. *Drugs*, *67*(3), 351–368.

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology*, *7*(10), e1002195.

Galperin, M. Y., & Koonin, E. V. (2004). "Conserved hypothetical" proteins: prioritization of targets for experimental study. *Nucleic Acids Research*, *32*(18), 5452–5463.

Gibson, K. E., Barnett, M. J., Toman, C. J., Long, S. R., & Walker, G. C. (2007). The symbiosis regulator CbrA modulates a complex regulatory network affecting the flagellar apparatus and cell envelope proteins. *Journal of Bacteriology*, *189*(9), 3591–3602.

Gibson, K. E., Campbell, G. R., Lloret, J., & Walker, G. C. (2006). CbrA Is a Stationary-Phase Regulator of Cell Surface Physiology and Legume Symbiosis in Sinorhizobium meliloti. *Journal of Bacteriology*, *188*(12), 4508–4521.

Hebbar, P., Berge, O., Heulin, T., & Singh, S. P. (1991). Bacterial antagonists of Sunflower (Helianthus annuus L.) fungal pathogens. *Plant and Soil*, *133*(1), 131–140.

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., … Bork, P. (2016). eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research*, *44*(D1), D286–93.

Jennings, D. B., Ehrenshaft, M., Pharr, D. M., & Williamson, J. D. (1998). Roles for mannitol and mannitol dehydrogenase in active oxygen-mediated plant defense. *Proceedings of the National Academy of Sciences*, *95*(25), 15129–15133.

Johnson, E. G., Sparks, J. P., Dzikovski, B., Crane, B. R., Gibson, D. M., & Loria, R. (2008). Plant-pathogenic Streptomyces species produce nitric oxide synthase-derived nitric oxide in response to host signals. *Chemistry & Biology*, *15*(1), 43–50.

Kets, E. P., Galinski, E. A., de Wit, M., de Bont, J. A., & Heipieper, H. J. (1996). Mannitol, a novel bacterial compatible solute in Pseudomonas putida S12. *Journal of Bacteriology*, *178*(23), 6665–6670.

Kishii, R., & Takei, M. (2009). Relationship between the expression of ompF and quinolone resistance in Escherichia coli. *Journal of Infection and Chemotherapy: Official Journal of the Japan Society of Chemotherapy*, *15*(6), 361–366.

Lalucat, J., Bennasar, A., Bosch, R., García-Valdés, E., & Palleroni, N. J. (2006). Biology of Pseudomonas stutzeri. *Microbiology and Molecular Biology Reviews: MMBR*, *70*(2), 510–547.

Lamont, I. L., Beare, P. A., Ochsner, U., Vasil, A. I., & Vasil, M. L. (2002). Siderophore-mediated signaling regulates virulence factor production in Pseudomonas aeruginosa. *Proceedings of the National Academy of Sciences*, *99*(10), 7072–7077.

Marcelletti, S., & Scortichini, M. (2014). Definition of Plant-Pathogenic Pseudomonas Genomospecies of the Pseudomonas syringae Complex Through Multiple Comparative Approaches. *Phytopathology*, *104*(12), 1274–1282.

Mastropaolo, M. D., Silby, M. W., Nicoll, J. S., & Levy, S. B. (2012). Novel genes involved in Pseudomonas fluorescens Pf0-1 motility and biofilm formation. *Applied and Environmental Microbiology*, *78*(12), 4318–4329.

Maurelli, A. T. (2007). Black holes, antivirulence genes, and gene inactivation in the evolution of bacterial pathogens. *FEMS Microbiology Letters*, *267*(1), 1–8.

Medina, M., & Sachs, J. L. (2010). Symbiont genomics, our new tangled bank. *Genomics*, *95*(3), 129–137.

Merhej, V., Georgiades, K., & Raoult, D. (2013). Postgenomic analysis of bacterial pathogens repertoire reveals genome reduction rather than virulence factors. *Briefings in Functional Genomics*, *12*(4), 291–304.

Meziane, H., VAN DER Sluis, I., VAN Loon, L. C., Höfte, M., & Bakker, P. A. H. M. (2005). Determinants of Pseudomonas putida WCS358 involved in inducing systemic resistance in plants. *Molecular Plant Pathology*, *6*(2), 177–185.

Morris, C. E., Sands, D. C., Vinatzer, B. A., Glaux, C., Guilbaud, C., Buffière, A., … Thompson, B. M. (2008). The life history of the plant pathogen Pseudomonas syringae is linked to the water cycle. *The ISME Journal*, *2*(3), 321–334.

Newell, P. D., Yoshioka, S., Hvorecny, K. L., Monds, R. D., & O'Toole, G. A. (2011). Systematic analysis of diguanylate cyclases that promote biofilm formation by Pseudomonas fluorescens Pf0-1. *Journal of Bacteriology*, *193*(18), 4685–4698.

O'Brien, H. E., Thakur, S., & Guttman, D. S. (2011). Evolution of plant pathogenesis in Pseudomonas syringae: a genomics perspective. *Annual Review of Phytopathology*, *49*, 269–289.

Oh, H.-S., H.-S., O., Kvitko, B. H., Morello, J. E., & Collmer, A. (2007). Pseudomonas syringae Lytic Transglycosylases Coregulated with the Type III Secretion System Contribute to the Translocation of Effector Proteins into Plant Cells. *Journal of Bacteriology*, *189*(22), 8277–8289.

Oh, H.-S., Park, D. H., & Collmer, A. (2010). Components of the Pseudomonas syringae type III secretion system can suppress and may elicit plant innate immunity. *Molecular Plant-Microbe Interactions: MPMI*, *23*(6), 727–739.

Patel, H. K., Matiuzzo, M., Bertani, I., Bigirimana, V. de P., Ash, G. J., Höfte, M., & Venturi, V. (2014). Identification of virulence associated loci in the emerging broad host range plant pathogen Pseudomonas fuscovaginae. *BMC Microbiology*, *14*, 274.

Preston, G. M. (2004). Plant perceptions of plant growth-promoting Pseudomonas. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *359*(1446), 907–918.

Rajasekaran, M. B., Nilapwar, S., Andrews, S. C., & Watson, K. A. (2010). EfeO-cupredoxins: major new members of the cupredoxin superfamily with roles in bacterial iron transport. *Biometals: An International Journal on the Role of Metal Ions in Biology, Biochemistry, and Medicine*, *23*(1), 1–17.

Ravooru, N., Ganji, S., Sathyanarayanan, N., & Nagendra, H. G. (2014). Insilico analysis of hypothetical proteins unveils putative metabolic pathways and essential genes in Leishmania donovani. *Frontiers in Genetics*, *5*, 291.

Robleto, E. A., Lopez-Hernandez, I., Silby, M. W., & Levy, S. B. (2003). Genetic Analysis of the AdnA Regulon in Pseudomonas fluorescens: Nonessential Role of Flagella in Adhesion to Sand and Biofilm Formation. *Journal of Bacteriology*, *185*(2), 453–460.

Römling, U., Galperin, M. Y., & Gomelsky, M. (2013). Cyclic di-GMP: the first 25 years of a universal bacterial second messenger. *Microbiology and Molecular Biology Reviews: MMBR*, *77*(1), 1–52.

Ruiz, J. A., Bernar, E. M., & Jung, K. (2015). Production of siderophores increases resistance to fusaric acid in Pseudomonas protegens Pf-5. *PloS One*, *10*(1), e0117040.

Silby, M. W., Winstanley, C., Godfrey, S. A. C., Levy, S. B., & Jackson, R. W. (2011). Pseudomonas genomes: diverse and adaptable. *FEMS Microbiology Reviews*, *35*(4), 652–680.

Soscia, C., Hachani, A., Bernadac, A., Filloux, A., & Bleves, S. (2007). Cross Talk between Type III Secretion and Flagellar Assembly Systems in Pseudomonas aeruginosa. *Journal of Bacteriology*, *189*(8), 3124–3132.

Stevanin, T. M., Read, R. C., & Poole, R. K. (2007). The hmp gene encoding the NO-inducible flavohaemoglobin in Escherichia coli confers a protective advantage in resisting killing within macrophages, but not in vitro: Links with swarming motility. *Gene*, *398*(1-2), 62–68.

Sudhamsu, J., & Crane, B. R. (2009). Bacterial nitric oxide synthases: what are they good for? *Trends in Microbiology*, *17*(5), 212–218.

Taguchi, F., Suzuki, T., Inagaki, Y., Toyoda, K., Shiraishi, T., & Ichinose, Y. (2009). The Siderophore Pyoverdine of Pseudomonas syringae pv. tabaci 6605 Is an Intrinsic Virulence Factor in Host Tobacco Infection. *Journal of Bacteriology*, *192*(1), 117–126.

Torres, M. A. (2006). Reactive Oxygen Species Signaling in Response to Pathogens. *Plant Physiology*, *141*(2), 373–378.

Tremaroli, V., Fedi, S., Tamburini, S., Viti, C., Tatti, E., Ceri, H., … Zannoni, D. (2011). A histidine-kinase cheA gene of Pseudomonas pseudoalcaligens KF707 not only has a key role in chemotaxis but also affects biofilm formation and cell metabolism. *Biofouling*, *27*(1), 33–46.

Veal, E. A., Toone, W. M., Jones, N., & Morgan, B. A. (2002). Distinct roles for glutathione S-transferases in the oxidative stress response in Schizosaccharomyces pombe. *The Journal of Biological Chemistry*, *277*(38), 35523–35531.

Wahyudi, A. T., Astuti, R. I., & Giyanto. (2011). Screening of Pseudomonas sp. Isolated from Rhizosphere of Soybean Plant as Plant Growth Promoter and Biocontrol Agent. *American Journal of Agricultural and Biological Science*, *6*.

Wheeler, N. E., Barquist, L., Kingsley, R. A., & Gardner, P. P. (2016). A profile-based method for identifying functional divergence of orthologous genes in bacterial genomes. *Bioinformatics* , *32*(23), 3566–3574.

Yeung, A. T. Y., Bains, M., & Hancock, R. E. W. (2011). The sensor kinase CbrA is a global regulator that modulates metabolism, virulence, and antibiotic resistance in Pseudomonas aeruginosa. *Journal of Bacteriology*, *193*(4), 918–931.

Yu, M., Xu, Y., Xu, T., Wang, B., Sheng, A., & Zhang, X.-H. (2015). WcaJ, the initiating enzyme for colanic acid synthesis, is required for lipopolysaccharide production, biofilm formation and virulence in Edwardsiella tarda. *Aquaculture* , *437*, 287–291.

Zumaquero, A., Macho, A. P., Rufián, J. S., & Beuzón, C. R. (2010). Analysis of the role of the type III effector inventory of Pseudomonas syringae pv. phaseolicola 1448a in interaction with the plant. *Journal of Bacteriology*, *192*(17), 4474–4488.

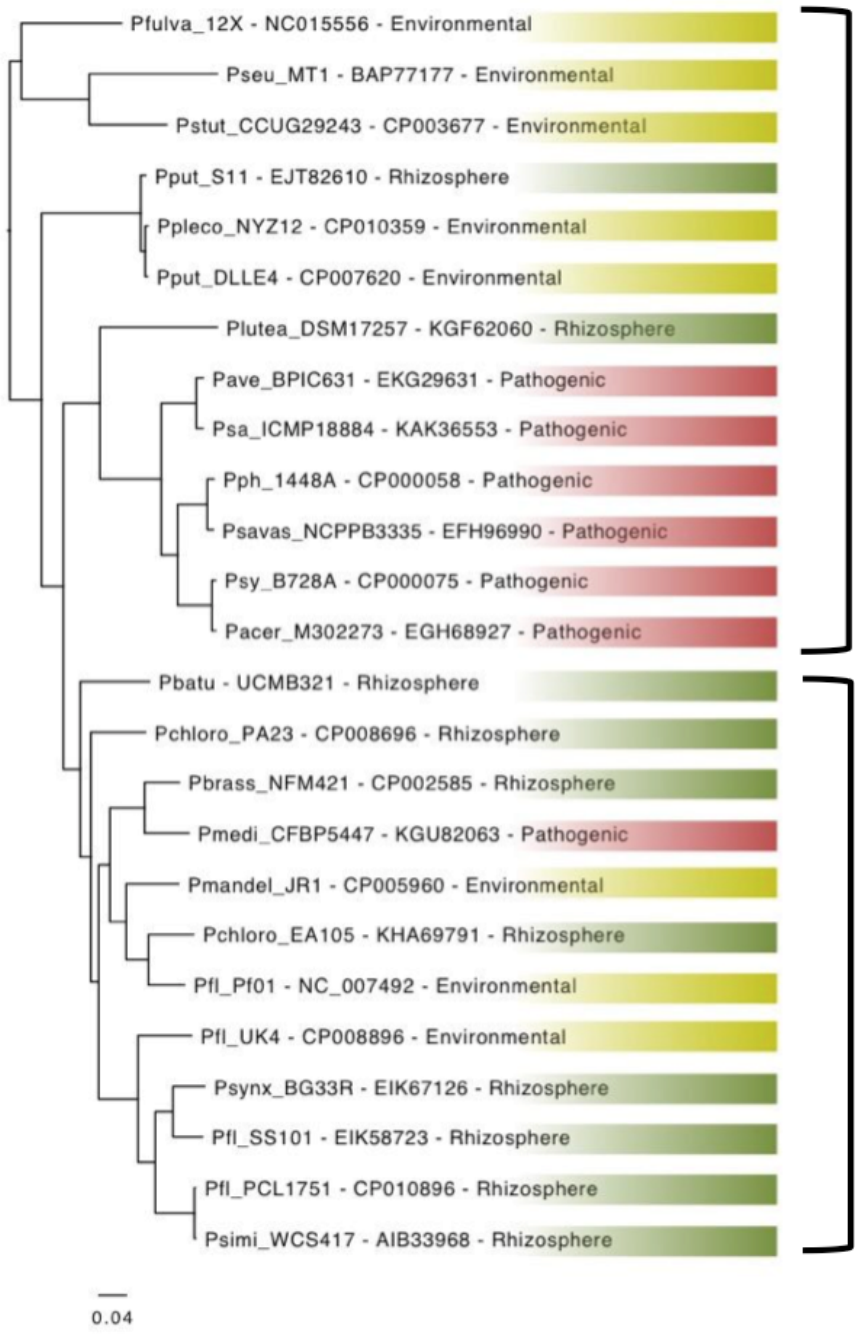
Supplementary Material

Supplementary Table 1 | *Pseudomonas* strains used to derive bitscores

Accession	Pathovar
Pathogenic	
EGH68927	Pseudomonas syringae pv. aceris str. M302273
EKG29631	Pseudomonas avellanae BPIC 631
KGS10723	Pseudomonas coronafaciens
KGU82063	Pseudomonas mediterranea CFBP 5447
CP000058	Pseudomonas syringae pv. phaseolicola 1448A
KAK36553	Pseudomonas syringae pv. actinidiae ICMP 18884
EFH96990	Pseudomonas savastanoi pv. savastanoi NCPPB 3335
CP000075	Pseudomonas syringae pv. syringae B728a
AE016853	Pseudomonas syringae pv. tomato str. DC3000
Rhizosphere associated	
KIH80415	Pseudomonas batumici
CP002585	Pseudomonas brassicacearum subsp. brassicacearum NFM421
KHA69791	Pseudomonas chlororaphis
CP008696	Pseudomonas chlororaphis strain PA23
CP010896	Pseudomonas fluorescens strain PCL1751
EIK58723	Pseudomonas fluorescens SS101
KGF62060	Pseudomonas lutea
EJT82610	Pseudomonas putida S11
AIB33968	Pseudomonas simiae
EIK67126	Pseudomonas synxantha BG33R
Environmental	
NC_007492	Pseudomonas fluorescens Pf0-1 chromosome
CP008896	Pseudomonas fluorescens strain UK4
NC_015556	Pseudomonas fulva 12-X chromosome
CP005960	Pseudomonas mandelii JR-1
CP010359	Pseudomonas plecoglossicida strain NyZ12
CP007620	Pseudomonas putida strain DLL-E4
BAP77177	Pseudomonas sp. MT-1
CP003677	Pseudomonas stutzeri CCUG 29243

Supplementary Table 2 | Top candidate genes identified

EggNOG model	*P. fluorescens* gene representative	Gene function	DBS (pathogen-nonpathogen)	DBS empirical P-value	KS statistic	KS P-value
ENOG410QR3U	Pfl01_2295	Diguanylate cyclase	147.2	0.006	0.67	0.008
ENOG410R1V2	Pfl01_2873	Periplasmic iron transport protein EfeO	144.5	0.006	1.00	0.000
ENOG410R78U	Pfl01_2182	Oligopeptide ABC transporter OppA	121.8	0.007	0.89	0.000
ENOG410QQIS	Pfl01_2897	Putative autotransporter of pectin lyase	112.9	0.008	0.80	0.001
ENOG410R96I	Pfl01_5484	HemY-like protein	112.35	0.008	0.89	0.000
ENOG410R1KB	Pfl01_2095	Hypothetical protein	103	0.008	0.88	0.001
ENOG410R9IN	Pfl01_0859	Lipoprotein A export system protein LptA	96.2	0.009	0.89	0.000
ENOG410QPVC	Pfl01_1777	Outer membrane fibronectin binding protein OmpF	89	0.009	0.61	0.023
ENOG410RA8U	Pfl01_5038	Putative membrane protein	76.7	0.011	0.81	0.002
ENOG410RAYZ	Pfl01_4924	Putative exonuclease	70.6	0.011	0.89	0.000
ENOG410R950	Pfl01_1538	Flagellar assembly protein FliH	68.2	0.012	0.78	0.002
ENOG410RA8Z	Pfl01_0583	Putative GCN5-like N-acetyltransferase	66.95	0.012	0.89	0.000
ENOG410QWN6	Pfl01_1032	Putative lipoprotein	64.3	0.013	0.73	0.011
ENOG410QHZG	Pfl01_2183	3 membrane-bound lytic murein transglycosylase D	62.6	0.013	0.78	0.001
ENOG410R2VH	(Psefu_2796)	LmbE family protein	61.85	0.013	0.61	0.043
ENOG410QJ6E	Pfl01_3829	Extracellular polysaccharide biosynthesis protein WcaJ	61.05	0.014	0.67	0.007
ENOG410QM5W	Pfl01_0786	Hypothetical protein	59	0.014	0.60	0.016
ENOG410RA8K	Pfl01_4535	L-glutamate/L-aspartate binding protein GltI	58.5	0.014	0.89	0.000
ENOG410QSW6	Pfl01_3622	Metal transport related protein	48.5	0.016	0.85	0.003
ENOG410QRC2	Pfl01_2834	Phosphoglycerate/bisphosphoglycerate mutase	47.95	0.016	0.75	0.004
ENOG410QJ0R	Pfl01_2636	Putative mannitol 2-dehydrogenase	46.6	0.017	0.78	0.002
ENOG410QPRG	Pfl01_2097	Non-homologous end joining protein LigD	46.45	0.017	0.67	0.005
ENOG410QTZ7	Pfl01_4329	GCN5-like N-acetyltransferase	46.2	0.017	0.86	0.004
ENOG410QYAE	Pfl01_1661	TetR family transcriptional regulator	44.3	0.017	0.79	0.005
ENOG410QSYD	Pfl01_4426	Glutathione S-transferase like protein	43.3	0.018	0.89	0.001
ENOG410R3KT	Pfl01_0814	Outer membrane porin	42.75	0.018	0.75	0.004
ENOG410QVZX	Pfl01_1852	Putative aminotransferase PvdN	40.7	0.019	1.00	0.000
ENOG410QUV1	Pfl01_4914	Putative exported protein	40.5	0.019	0.89	0.000
ENOG410RA00	Pfl01_2608	Major facilitator superfamily transporter	39.2	0.020	0.67	0.015
ENOG410QIG1	Pfl01_4086	Diguanylate cyclase/phosphodiesterase RbdA	-38.1	0.021	0.82	0.000
ENOG410QNPT	Pfl01_3125	Iron uptake protein FecR	-38.4	0.021	0.69	0.009
ENOG410QK68	Pfl01_0939	Fuscaric acid resistance protein	-39.05	0.020	0.86	0.000
ENOG410QRT5	Pfl01_4022	methyl accepting chemotaxis protein	-39.65	0.020	0.81	0.001
ENOG410QX18	Pfl01_0909	LclR family transcriptional regulator	-39.8	0.020	0.81	0.001
ENOG410QJN6	Pfl01_2361	LysR family transcriptional regulator	-40.95	0.019	0.83	0.000
ENOG410QIH8	Pfl01_4652	Flavohemoprotein Hmp	-41.3	0.019	0.94	0.001
ENOG410QPSB	Pfl01_1300	Putative membrane protein	-41.9	0.018	0.76	0.003
ENOG410QQDR	Pfl01_0066	D-methionine ABC transporter	-42.6	0.018	0.83	0.001
ENOG410QM4H	Pfl01_2693	Sodium hydrogen antiporter	-46.3	0.017	0.81	0.002
ENOG410RAFM	Pfl01_4807	Two-component system sensor kinase CbrA	-50	0.016	0.65	0.014
ENOG410QIBS	Pfl01_3863	FAD:protein FMN transferase AbpE	-53.1	0.015	0.89	0.000
ENOG410R80A	Pfl01_2672	DNA helicase	-56	0.015	0.76	0.002
ENOG410QJGJ	Pfl01_2909	CdaR family transcriptional regulator	-62.45	0.013	0.76	0.002
ENOG410RA58	Pfl01_0797	Flavodoxin nitric oxide synthase	-67.7	0.012	0.69	0.009
ENOG410QJXA	Pfl01_3630	Glycerol transporter	-68.8	0.012	0.76	0.003
ENOG410RB3U	Pfl01_1833	Putative aerotaxis receptor	-72.6	0.011	0.71	0.005
ENOG410R8YE	Pfl01_5304	Signal transduction histidine kinase CheA	-73.6	0.011	0.58	0.047
ENOG410R9RW	Pfl01_1782	Serine threonine-protein kinase	-86.7	0.009	0.76	0.002
ENOG410RAGW	Pfl01_3260	Putative ABC transporter, periplasmic polyamine binding protein	-129.95	0.007	0.83	0.000
ENOG410QJAF	(Psefu_1476)	Error-prone lesion bypass DNA polymerase V (UmuC)	-162.65	0.005	0.60	0.039
ENOG410QHSC	Pfl01_3975	Multi sensor hybrid histidine kinase	-171.3	0.004	0.65	0.007



Supplementary Figure 1 | Whole genome phylogeny

RAxML tree constructed using 3.46Mb Mugsy whole genome alignment (2.97M variable positions) with GTRGAMMA substitution model. Bootstrapping IP. Taxon highlight colour according to niche of origin (red: pathogenic, green: rhizosphere, yellow: environmental). The two groups used for the control comparison are indicated with brackets.

Chapter Seven | Draft: Phylogenetic conservation of metabolic capabilities in *Staphylococcus* species

Preface

This chapter departs from the development of the delta-bitscore method, and instead looks at another rich source of data on bacterial adaptation: the phenotype array. For this study we used the Biolog GEN III MicroPlate™ to characterise the breadth of carbon sources and stress conditions that Staphylococcus isolates could grow on/in, and to assess whether these growth characteristics showed a strong phylogenetic signal. The Biolog GEN III MicroPlateTM has been designed for the identification of a broad range of Gram-negative and Gram-positive bacteria using a standard array of 94 biochemical tests. Because these tests had been carefully selected for their ability to distinguish bacteria to the species level, we were curious to see which tests were informative of species, and these tests could accurately identify *Staphylococcus* species.

We identified some distinguishing features in the metabolic profiles of the different species included in our study. We observed that *S. aureus* and *S. lugdunensis*, which are associated with more aggressive forms of infection also showed some of the most vigorous growth and metabolic flexibility of the isolates tested, traits that may give them a competitive advantage during infection. Across a broad range of substrates, we observed as much variation in the ability to utilize a substrate within species as we did between closely related species. This overlap and variability in metabolic profiles was reflected in the high misclassification rate of the Biolog GEN III classification system.

This study not only assesses differences in the growth characteristics of staphylococci but also assesses the practical limitations to performing Biolog phenotype arrays on a large scale. We found that strains in our study showed very different growth characteristics in the negative control well of the GEN III plate. While growth in this well is unexpected, we observed it in approximately one third of our samples, greatly complicating the analysis and interpretation of our findings. I present a potential solution to analysing plates that show growth on the negative control well, in order to facilitate their inclusion in comparative studies, by scaling growth parameters relative to both the positive and negative control wells.

Contributions

Gemma Langridge and colleagues performed Biolog experiments and sequencing, I performed the analysis and wrote the paper.

Phylogenetic conservation of metabolic capabilities in *Staphylococcus* species

Nicole E. Wheeler[1], Rebecca Clifford[2], Emma J. Meader[2], John Wain[2], Lisa C. Crossman[2,4,5], Claire Hill[2], Gemma C. Langridge[2]

1. School of Biological Sciences, University of Canterbury, Christchurch, New Zealand.
2. Medical Microbiology Research Laboratory, University of East Anglia, Norwich, NR4 7UQ, UK.
3. Norfolk & Norwich University Hospital, Colney Lane, Norwich, NR4 7UY, UK.
4. School of Biological Sciences, University of East Anglia, Norwich, NR4 7UQ, UK.
5. SequenceAnalysis.co.uk, NRP Innovation Centre, Norwich Research Park, Norwich, NR4 7JG, UK.

Abstract

Staphylococci are common human skin commensals, capable of colonising diverse niches in the human body, however they can also be clinically important pathogens. This diversity in site colonisation and pathogenic potential has been linked with differences in metabolic capacity. Here we use the Biolog GEN III MicroPlate™ to characterise the diversity in carbohydrate utilization capacities and tolerance to environmental pressures found in 237 *Staphylococcus* isolates. We outline some methodological considerations for the scaling of Biolog phenotype array analyses to large numbers of strains, and observe different levels of phylogenetic conservatism across a range of carbon substrate utilization profiles. Across the 94 biochemical tests employed by the Biolog GEN III array, we were unable to identify growth characteristics that were unique to individual species, however we did observe faster growth and greater metabolic flexibility in more clinically aggressive species of staphylococci.

Introduction

Staphylococci are common skin commensals (Bannerman, 2003), however they are also major contributors to the incidence of nosocomial infections worldwide (Becker et al., 2014; Naber, 2009; Tan et al., 2006). *Staphylococcus aureus* is associated with more aggressive infections ranging from skin infections to invasive disease (Lowy, 1998), while coagulase negative staphylococci (CoNS) tend to have lower virulence and are more commonly associated with infection resulting from contamination of indwelling medical devices (Rogers et al., 2009). Within the staphylococci, individual species are differentially associated with specific infections and colonisation sites on the body (Becker et al., 2014), suggesting niche

adaptation linked with differences in their ability to both obtain site-specific nutrients and tolerate a range of stressors. We sought to systematically characterise differences in the ability to utilise a range of carbon sources and tolerate stress conditions across *Staphylococcus* species, in order to better understand the clinical implications of species classifications. We were interested not only in inter-species differences in metabolism, but also in intra-species differences suggestive of niche adaptation.

The shared genetic ancestry of a species implies many similarities in the niches it can inhabit. However these similarities can be rapidly altered due to gene loss and gain events, which can change the ecological niche of a bacterium (Martiny et al., 2013). Comparative genomics analyses have found that a large proportion of the difference in gene presence/absence within a clade occur among closely related species (Kettler et al., 2007), representing recent acquisition and loss events that may later be reversed or lost from the population. For example, transporters and enzymes for sugar and amino acid metabolism are frequently acquired through HGT or duplication, allowing rapid shifts in metabolic capabilities (Makarova et al., 2006). Many of these differences in the presence and absence of genes, as well as gene degradation appear to be more closely linked to nutrient availability in the niche a bacterium occupies than phylogeny, resulting in convergence on similar metabolic strategies that provide a survival benefit in a given niche (Kettler et al., 2007; Nuccio & Baumler, 2014). This potential for convergence of metabolic traits prompted us to investigate whether species boundaries in *Staphylococcus* were reflected in strong differences in growth characteristics on a range of substrates, or whether the growth characteristics of individual strains showed patterns of similarity that ran counter to phylogeny.

We found that unsupervised clustering of isolates by growth characteristics caused genetically similar isolates to group together, but did not result in clear separation of species. There was a particular overlap in the growth characteristics of *S. epidermidis* and *S. capitis*, to the point where their growth characteristics using the biochemical tests provided by the Biolog phenotype array were indistinguishable. *S. aureus* and *S. lugdunensis* demonstrated the fastest growth, and the broadest ability to utilise a range of carbon sources, possibly reflective of their more aggressive mode of infection (Becker et al., 2014). We noted broad similarities in the response of *Staphylococcus* species to a range of stress conditions, with intra-species variability often rivaling inter-species variability.

During our analysis, we identified methodological challenges in scaling Biolog phenotype arrays to hundreds of bacterial samples, namely the inconsistency of growth curves between replicates for a small subset of wells, and growth of an subset of the samples in the negative control well of the plate. These methodological challenges are discussed, and we provide our recommendations for researchers hoping to use this technology in a high-throughput manner in the future.

Method

Clinical samples were obtained from the Norfolk and Norwich University Hospital (NNUH), in order to assemble a large collection of staphylococci. We have supplemented these with strains isolated from healthy people and animals, performed whole genome sequencing and Biolog phenotype arrays. We aimed to assess whether genetic relationships between these *Staphylococcus* strains are reflected in metabolic similarities. We investigated which metabolic capabilities were broadly conserved, and which showed weak phylogenetic structure, to assess which functions were more fundamental to the functioning of a species and which were more niche-specific.

Strain collection

Strains were isolated from clinical specimens at the Norfolk and Norwich University Hospital (NNUH), and from skin swabs of 10 healthy volunteers. Consecutive isolates were obtained from all clinical specimens in 3 months in 2013, and from samples considered significant by the clinical microbiology lab from 2013-2016. Animal isolates (dogs, cat, sheep, calf and alpaca) were also obtained from skin swabs. All isolates were cultured on CLED agar (Oxoid, UK) and incubated at 37^{o}C overnight. Single colonies were identified using MALDI-TOF MS (Bruker). All coagulase negative staphylococci were frozen and deposited in the NNUH Biorepository.

Sequencing

For each strain, DNA was extracted from 1 mL of overnight iso-sensitest broth (Oxoid) culture. After centrifugation of the 1 mL, the cell pellet was resuspended in lysis buffer (Qiagen) and then transferred to 2 mL lysis matrix B tubes (MPBio) for bead beating (15mins at 30 Hz using Tissuelyser II, Qiagen) with 2 µl of RNAse A (Qiagen). Extraction was performed using the standard QiaCube HT protocol (Qiagen) with incubation for 30 mins at 65^{o}C after proteinase K addition, then elution into 70 µL of 10 nM Tris-HCL. Sequence libraries were prepared using the Nextera XT DNA Library Prep Protocol (Illumina). The

average fragment length for each library was checked using a Tapestation (Agilent) and quantified with Picogreen to allow manual normalization to 4 nM for sequencing. Samples were sequenced on the Illumina NextSeq, with a loading concentration of 1.8 pM. 16S sequences were identified on DNA sequence assemblies by solely running Barrnap (http://www.vicbioinformatics.com/software.barrnap.shtml) as part of Prokka (Seemann, 2014) in parallel. Sequences were parsed from gff files with a Python (2.7.10) script, identifying on which strand the 16S sequence was present and writing separate fasta sequence files for each strain. All fasta sequences were concatenated and aligned with MAFFT (Katoh & Standley, 2013), the alignment was examined and curated in Python to remove ragged ends and ensure full length 16S gene was included. The average read coverage of each gene across the samples was 51.27%. Reference 16S sequences for *Staphylococcus* species were included in the alignment for species identification. A tree was created with FastTreeMP (Price, Dehal, & Arkin, 2010). Species classifications were made by grouping strains with the most closely related *Staphylococcus* reference isolate.

Phenotype microarrays

Metabolic capabilities were assayed using Biolog GEN III phenotype microarray plates (Biolog). The Biolog GEN III MicroPlate™ has been designed for the identification of a broad range of Gram-negative and Gram-positive bacteria using a standard array of 94 biochemical tests, as well as positive and negative control wells (Biolog, Inc, 2013). Growth in the wells is measured by the reduction of a tetrazolium dye, which changes colour as cells metabolise a biochemical substrate and produce energy, presumably NADH (Bochner et al., 2011). The plate contains 71 carbon source utilization assays and 23 chemical sensitivity assays. Carbon source utilization wells will have a nearly universal growth medium, with all substrates required for growth, but only a single carbon source provided. If a cell has the required transport system and catabolic pathway to utilise that carbon source, it will catabolise the chemical and produce NADH in the process, causing a colour change in the tetrazolium dye, resulting in a purple colour (Bochner, 2009). Chemical sensitivity assays carry all substrates required for growth, plus a potentially inhibitory substance. If cells grow, they will produce a colour change, but if they do not then the dye will remain colourless.

237 *Staphylococcus* samples were inoculated into BIOLOG GEN III plates to assay their growth properties over 22 hours. Protocol A was used for all isolates (Biolog, Inc, 2013). For each sample, a single colony was isolated and emulsified into inoculating fluid A according to manufacturer instructions. 100 µL of cell suspension was inoculated into each well of the

Biolog GEN III plate. Several strains were replicated to test for consistency. Colour intensity was recorded every 15 minutes by an OmniLog® PM system. For those species that give a false positive in the negative control well, the manufacturers of Biolog plates recommend testing the sample using protocol B instead (Biolog, Inc, 2013). Protocol B simply provides less media content for microorganisms to grow on, thus reducing colour change in the negative control well (Sandle et al., 2013). Staphylococci were not among the species that are usually recommended for this protocol, however a substantial proportion (77/237) of isolates met the criterion of giving a false positive in the negative control well (maximum intensity > 100 OmniLog® units, according to (Vaas, et al., 2012)). For our analysis, we wanted to be able to directly compare results across plates, so using one protocol for approximately one third of the isolates and another for the other two thirds would have been inappropriate. Instead, data were scaled to account for growth in the negative control well, as outlined in the following section.

Curve parameter estimation

It is typically advised that intensity values for the negative control well be subtracted from the other wells at each time point to subtract experimental noise from the measurements (Vaas et al., 2012), however because some bacteria appeared to be respiring in the control well, we decided it was inappropriate to simply subtract the control well from each of the other wells, as there was an upper limit on intensity of dye colour that was measured during our study (~300 OmniLog® units). As a result of this upper limit to the potential for colour change, subtraction of negative control well readings would place an lower limit on the growth parameters achievable by the high respirers compared to those that did not respire in the control well. To remedy this, parameters in each well were calculated as a proportion of those shown in the negative and positive control wells according to the formula:

$$x_e = (x_i - x_n) / (x_p - x_n)$$

where x_e is the “substrate utilization efficiency” parameter being calculated for a well of interest, x_i is the area under the growth curve (AUC) for that well, x_n is the AUC for the negative control well and x_p is the AUC for the positive control well. In other words, the metric is calculated as the growth observed by a sample on a given substrate, minus the growth of that sample without supplementation, as a proportion of growth under rich media conditions minus growth with no supplementation. For carbon sources, this metric is expected to fall between zero (equivalent to growth in the negative control well) and one

(equivalent to growth in the positive control well), but could occasionally fall below zero in the case where growth by the sample is impaired by supplementation, or above one where the sample grows better on a sole carbon source than on a mixture. For inhibitory substances, this metric is expected to fall below zero in cases where respiration is occurring in the negative control well and inhibited in the test well.

The R package grofit (Kahm et al., 2010) was used to fit curve parameter estimates with a smoothed splines approach. Curve parameter estimates were calculated using five bootstrap samples to improve the robustness of the estimates. AUC was expected to be the most informative parameter, as it correlates with the other parameters and captures the overall growth pattern best. As such, clustering was performed using AUC values adjusted using the equation above.

Cluster analysis

For cluster analysis, missing values that could not be calculated using the grofit package using the available data were imputed using a k nearest neighbours approach from the R package “impute” (Hastie et al., 2016) (228 out of a total 26400 AUC values were imputed). Principal component analysis (PCA) was performed using R (R Core Team, 2016), to identify any overall structure in the data. Next, the R package “pvclust” (Suzuki & Shimodaira, 2006) was used to perform hierarchical clustering via multiscale bootstrap resampling using a euclidean distance matrix computed from the AUC data.

Results

Growth parameters were largely consistent across replicates

To assess the consistency of growth in each well across replicates, curve parameter estimates between replicates of the same strain were compared. Parameter estimates for most samples did not vary considerably, however there were some notable exceptions, with differences in some utilization profiles reflecting efficient utilization of a substrate in one replicate and failure to utilise the substrate in the other. See Supplementary Table 1 and Supplementary Figure 1 for summary statistics, and Supplementary Figure 2 for an example of the consistency of growth curves across four replicates of the same sample.

A large proportion of samples showed activity in the negative control well

We noted that a number of strains were showing signs of growth in the negative control well (77/237 strains showed maximum intensity values >100 in control wells, which was the cutoff used to identify respiration in (Vaas et al., 2012)) (see Supplementary Figure 3 for the distribution of intensity values observed in the negative control well). Other authors have also noted growth in the negative control wells of Biolog plates (Vaas et al., 2012), indicating that this is likely to be a recurring issue that should be addressed using a standardised approach. Growth in the negative control wells was reproducible based on replicate data, while there were clear signs of suppressed respiration in some of the inhibitory substance wells, indicating that colour change in the negative control well was likely due to bacterial activity (Supplementary Figure 4). This unexpected observation prompted a change to the analysis method used, in which we scaled the aurea under the growth curve to reflect the proportion of activity observed compared to both the positive and negative control wells. This approach was expected to predominantly produce scaled AUC estimates that range between zero and one with a small number of exceptions, and this appears to have been the case. Notably, there is no clear separation between positive and negative signs of growth in the wells (Figure 1), indicating that a lot of subtlety is lost when Biolog substrate utilization data is converted to binary classifications of utilization and non-utilization.

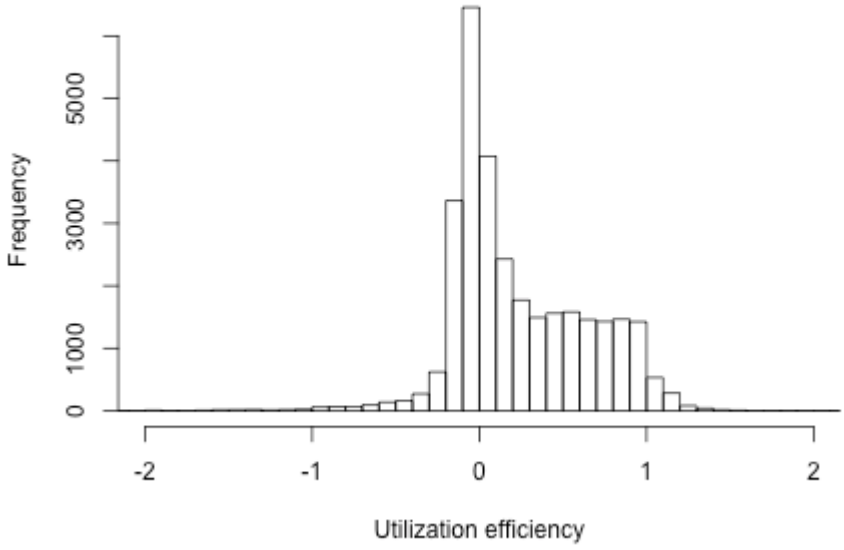


Figure 1 | An illustration of the utilization efficiencies calculated for all wells and samples

There is a peak around zero indicating no growth relative to the negative control, a small number of wells showing less growth than the negative control, and a smaller number of

wells showing better growth than the positive control wells, with most values falling between one and zero.

Incorrect identification of samples using Biolog GEN III classification criteria

Biolog GEN III plates are designed to distinguish bacterial species based on their metabolic characteristics, so we assessed the accuracy of the species classifications made using the species classification criteria provided by Biolog, compared to the similarity of 16S sequences of our samples to reference *Staphylococcus* species isolates. 28 strains (~10%) confirmed as belonging to *Staphylococcus* were identified as being from other genera using the Biolog GEN III species identification criteria. This indicates that identifying species based on metabolic data obtained from the GEN III plate is not an accurate approach. Of the strains that were misidentified on the genus level, utilization efficiencies do not look markedly different to those of close relatives, suggesting that a slight deviation in utilization across these wells can result in assignment to a very different phylogenetic group.

We hypothesised that the misidentified samples may have been those that grew in the negative control wells, however the distribution of AUC values for the negative control wells was not significantly different in misidentified samples compared to the full set of samples ($P = 0.525$, Mann Whitney U test). In addition, species classifications within the staphylococci were not consistent between Biolog and 16S approaches, however both approaches have been documented as mis-classifying bacteria at the species level before (Wragg et al., 2014). For further analyses, we used the 16S sequence data to group species for the comparison of their metabolic profiles, rather than using the Biolog GEN III species classification, due to the fact that 16S classification is more accurate (Wragg et al., 2014), and the fact that the Biolog criteria misidentified the genus of a large proportion of our samples.

Clustering of strains based on their metabolism reveals limited phylogenetic signal

We performed PCA on the substrate utilization efficiency values calculated for each strain in order to visualise any overall structure in the data consistent with species classifications. We observed clustering of strains according to species, however these clusters overlapped to such an extent that discrimination between species was not possible (Supplementary Figure 5). *S. epidermidis, S. capitis* and *S. hominis* clustered together, while *S. aureus, S. haemolyticus* and *S. lugdunensis* formed a separate cluster, which showed greater dispersion. *S. warneri* samples were highly dispersed and overlapped the two major

groupings, showing greater similarity to *S. haemolyticus* than the *epidermidis/capitis* grouping. Hierarchical clustering of strains confirmed these findings, with two major groupings identified in the data, and species interspersed within these two groups.

Phylogenetic conservation of metabolic profiles

Overall, differences in utilization of specific substrates were noticeable across species (Supplementary Figure 6). All strains grew comparably well in the positive control well, however *S. aureus, S. haemolyticus* and *S. lugdunensis* also grew in the negative control well. *S. aureus S. haemolyticus*, *S. lugdunensis* and *S. warneri* were also able to utilise a wider range of carbon sources for growth above the rate observed in the negative control well. In contrast, *S. capitis, epidermidis* and *hominis* showed less vigorous growth across wells, and were able to effectively utilise fewer substrates, however a handful of strains within these species showed substrate utilization profiles more akin to faster growing strains from other species.

We observed some growth characteristics that were consistent across species. All samples grew particularly well on α-D-glucose, followed by dextrin and pectin, then by sucrose, D-maltose and fructose. Malic acid, D-salicin, β-methyl-D-glucoside and bromo-succinic acid were all mildly inhibitory to growth. Species were also resilient to a range of inhibitory conditions. All species grew equally well on the three concentrations of salt tested (1%, 4% and 8% NaCl), but grew better at pH6 than at pH5. We observed resistance of approximately one in ten isolates to troleandomycin and lincomycin across all species except *S. aureus*, which was the least sampled of all species and contained no resistant isolates. In contrast we observed resistance of only a handful of isolates to fusidic acid, rifamycin, minocycline or vancomycin.

While some substrates showed consistent patterns of growth or no growth across species, others showed high inter-strain variability in utilization. Utilization efficiencies of D-fructose-6-P, D-glucose-6-P, D-serine, D-mannose, glycerol, D-gluconic acid, D-lactic acid and acetic acid were particularly variable across closely related strains. We noticed some complementarity in strains that did not utilise D-fructose-6-P, D-glucose-6-P as efficiently, but utilised D-mannose more efficiently. All three of these substrates can feed into early glycolysis (Balibar, Shen, & Tao, 2009), so this finding may represent differences in preferred carbohydrate sources or efficiencies in transporting the different substrates. Tolerance to D-serine, which can be found in a range of sites around the human body and is secreted at

high concentrations in urine (Roesch et al., 2003), has previously been observed in *S. saprophyticus*, but not in other *Staphylococcus* species (Sakinç et al., 2009), however here we see that at the concentration used for the Biolog GEN III plate, a broader collection of staphylococci exhibit tolerance.

Ultimately, we sought to identify substrate utilization patterns that could be used to distinguish the species in our study. The metabolic profiles of *S. epidermidis* and *S. capitis* were indistinguishable (Supplementary Figures 7 and 8), with as much variation in utilization efficiencies within the species as between them. *S. hominis* and *S. warneri* showed greater susceptibility to the inhibitory effects of D-serine, allowing them to be separated from other species. *S. hominis* could be distinguished by having similar utilization profiles for fucose, melibiose, gentiobiose, rhamnose, glucuronic acid and galacturonic acid as *S. haemolyticus, lugdunensis* and *warneri*, whilst having lower utilization values more akin to *S. epidermidis* and *S. capitis* for a range of other substrates. *S. aureus, haemolyticus* and *lugdunensis* showed the greatest ability to utilise GlcNAc and ManNAc, while other species showed mixed abilities and most of *S. epidermidis* and *S. capitis* failed to utilise these carbon sources. While trends could be identified between species, these were not sufficient to completely separate the species for classification purposes.

Discussion

In this study, we have collected a large volume of comparable phenotype arrays from a collection of coagulase negative staphylococci. We have found that while some phylogenetic structure exists in the metabolic profiles of these staphylococci, there is often as much variability in the ability to utilize a substrate within species as there is between them.

Methodological considerations

The study of large numbers of bacteria using high throughput phenotype arrays requires an understanding of the methodological issues that can arise when comparing so many samples. We have identified several methodological issues during our study which will require greater consideration in future high-throughput work. Firstly, while most replicates showed consistent curve parameters across most wells tested, there were a few notable outliers, with some replicates showing discrepancies in curve parameters spanning almost the full range of possible values. This generally indicates that in some cases one replicate would fail to grow in a well and another would grow well. This raises questions about the reliability of data from single replicate strains, as well as how to interpret replicate data for

the same strain. Our findings indicate the need for replicates for each sample to ensure ability to utilise each substrate is accurately assessed.

In addition to inconsistency across replicates, we observed growth of some of our staphylococci in the negative control well of the GEN III plate, a phenomenon that has been observed by other users as well (Vaas et al., 2012). For processing of individual samples, the correct solution to this problem would be to repeat the experiment using Protocol B provided by the manufacturers (Biolog, Inc, 2013), however when processing large numbers of samples, we decided that using different preparation protocols for one third of our samples would prevent fair comparison of samples. As a solution to this, we have scaled growth parameters in each well to represent a proportion of the difference in growth seen between the negative and positive control wells. This solution will be unique to Biolog plates that have both positive and negative control wells, however, and plates that only have a negative control well would need to be corrected differently, perhaps by taking a proportion relative to a simulated curve representing the fastest growth parameters achievable given the intensity range detectable using this approach.

Metabolism of Staphylococcus species

Different sites in the body show different levels of nutrient availability, as do different individuals (Krismer et al., 2014), and thus we may expect that the isolates from this study have originated from different body sites and thus have developed different preferences for carbon sources. A study of the core and pan genomes of *S. aureus* isolates found that of the genes conferring the pan-metabolism of this species, the largest group of predicted metabolic capabilities not shared by all isolates was carbohydrate metabolism genes, largely owing to the presence and absence of genes required for the metabolism of niche-specific carbon sources (Bosi et al., 2016). In *E. coli*, we see that pathogenic strains utilise additional sugars that are not used by commensal strains for energy, giving them a competitive advantage (Fabich et al., 2008), and it may be that the isolates we are investigating have developed similar metabolic specialties that allow them to establish long-term colonisation of a host and in some cases outcompete resident commensal bacteria to establish a clinical infection.

We observed an overall difference in the rate of growth of different *Staphylococcus* species on the Biolog plates. Generally speaking, *S. aureus, S. lugdunensis* and *S. haemolyticus* showed better growth across a range of carbon sources, despite comparable areas under

the growth curve for the positive control well. This finding is supported by the observation of Krismer et al (2014) of a shorter lag time for growth of *S. aureus* compared to *S. epidermidis* when grown in nutrient conditions intended to mimic the human nose. We observed a similar pattern across a range of growth substrates, leading to consistently high areas under the growth curve for *S. aureus* strains. Krismer et al. (2014) also noted large intra-species variation in the ability of CoNS to grow in the conditions they tested, a result which is echoed here for a wide range of potential carbon sources.

The classification criteria employed by the Biolog GEN III plate was unable to accurately distinguish species of staphylococci in our study. Inaccuracies in species classifications using Biolog GEN III classification criteria have been observed previously (Wragg et al., 2014), which together with our findings suggests there is limited utility in using information on the growth of bacteria on such a small collection of substrates in classifying bacteria. We saw a large overlap in substrate utilization efficiencies of *Staphylococcus* species, indicating that the array of carbohydrate sources and inhibitory substances in the GEN III plate are insufficient to classify samples to the species level.

Despite the generally different clinical presentation of *S. lugdunensis* compared to the *S. epidermidis*-like group (incl. *S. epidermidis, S. haemolyticus, S. capitis, S. hominis* and *S. warneri*) (Becker et al., 2014), the metabolic profiles of the species did not directly reflect these divisions. *S. lugdunensis* was grouped in the same major cluster as *S. aureus*, consistent with its clinical behaviour being more aggressive and closer to that of *S. aureus* than other CoNS (Babu & Oropello, 2011), however all isolates of *S. haemolyticus* from the study, as well as a handful of other *S. epidermidis*-like group CoNS, including the majority of *S. warneri* isolates, also fell into this group, which was characterised by its faster growth across most media conditions.

Concluding statements

We have investigated the value of the Biolog GEN III MicroPlate™ test panel in distinguishing *Staphylococcus* species and characterising their metabolic capacity. We found that it had limited accuracy in correctly identifying staphylococci to either the genus or species level. We did, however, find that the technology was able to show general differences in the rate of growth between more aggressive staphylococci such as *S. aureus* and *S. lugdunensis*, compared to strains with milder clinical presentations such as *S. epidermidis* and *S. capitis*. We observed high intra-species variation in the ability to utilise a

range of carbon sources, which likely contributed to much of the difficulty in correctly classifying the species, but which hints at metabolic differences which could provide clinically useful indicators of niche specialisation and ability to compete for nutrients with commensal bacteria in an infection context.

References

Babu, E., & Oropello, J. (2011). Staphylococcus lugdunensis: the coagulase-negative staphylococcus you don't want to ignore. *Expert Review of Anti-Infective Therapy*, *9*(10), 901–907.

Balibar, C. J., Shen, X., & Tao, J. (2009). The mevalonate pathway of Staphylococcus aureus. *Journal of Bacteriology*, *191*(3), 851–861.

Bannerman, T. L. (2003). Staphylococcus, Micrococcus, and other catalase-positive cocci that grow aerobically. *Manual of Clinical Microbiology*, *8*, 384–404.

Becker, K., Heilmann, C., & Peters, G. (2014). Coagulase-negative staphylococci. *Clinical Microbiology Reviews*, *27*(4), 870–926.

Biolog, Inc. (2013). *GEN III MicroPlate TM: Instructions for Use*. Retrieved from http://www.biolog.com/pdf/milit/00P%20185rA%20GEN%20III%20MicroPlate%20IFU%20Mar2008.pdf

Bochner, B. R. (2009). Global phenotypic characterization of bacteria. *FEMS Microbiology Reviews*, *33*(1), 191–205.

Bochner, B. R., Siri, M., Huang, R. H., Noble, S., Lei, X.-H., Clemons, P. A., & Wagner, B. K. (2011). Assay of the multiple energy-producing pathways of mammalian cells. *PLoS One*, *6*(3), e18147.

Bosi, E., Monk, J. M., Aziz, R. K., Fondi, M., Nizet, V., & Palsson, B. Ø. (2016). Comparative genome-scale modelling of Staphylococcus aureus strains identifies strain-specific metabolic capabilities linked to pathogenicity. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(26), E3801–9.

Fabich, A. J., Jones, S. A., Chowdhury, F. Z., Cernosek, A., Anderson, A., Smalley, D., … Conway, T. (2008). Comparison of Carbon Nutrition for Pathogenic and Commensal Escherichia coli Strains in the Mouse Intestine. *Infection and Immunity*, *76*(3), 1143–1152.

Hastie, T., Tibshirani, R., Narasimhan, B., & Chu, G. (2016). impute: impute: Imputation for microarray data.

Kahm, M., Hasenbrink, G., Lichtenberg-Fraté, H., Ludwig, J., Kschischo, M., & Others. (2010). grofit: fitting biological growth curves with R. *Journal of Statistical Software*, *33*(7), 1–21.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780.

Kettler, G. C., Martiny, A. C., Huang, K., Zucker, J., Coleman, M. L., Rodrigue, S., … Chisholm, S. W. (2007). Patterns and implications of gene gain and loss in the evolution of Prochlorococcus. *PLoS Genetics*, *3*(12), e231.

Krismer, B., Liebeke, M., Janek, D., Nega, M., Rautenberg, M., Hornig, G., … Peschel, A. (2014). Nutrient limitation governs Staphylococcus aureus metabolism and niche adaptation in the human nose. *PLoS Pathogens*, *10*(1), e1003862.

Lowy, F. D. (1998). Staphylococcus aureus infections. *The New England Journal of Medicine*, *339*(8), 520–532.

Makarova, K., Slesarev, A., Wolf, Y., Sorokin, A., Mirkin, B., Koonin, E., … Mills, D. (2006). Comparative genomics of the lactic acid bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(42), 15611–15616.

Martiny, A. C., Treseder, K., & Pusch, G. (2013). Phylogenetic conservatism of functional traits in microorganisms. *The ISME Journal*, *7*(4), 830–838.

Naber, C. K. (2009). Staphylococcus aureus bacteremia: epidemiology, pathophysiology, and management strategies. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, *48 Suppl 4*, S231–7.

Nuccio, S.-P., & Baumler, A. J. (2014). Comparative Analysis of Salmonella Genomes Identifies a Metabolic Network for Escalating Growth in the Inflamed Gut. *mBio*, *5*(2), e00929–14–e00929–14.

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PloS One*, *5*(3), e9490.

R Core Team. (2016). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Roesch, P. L., Redford, P., Batchelet, S., Moritz, R. L., Pellett, S., Haugen, B. J., … Welch, R. A. (2003). Uropathogenic Escherichia coli use d-serine deaminase to modulate infection of the murine urinary tract. *Molecular Microbiology*, *49*(1), 55–67.

Rogers, K. L., Fey, P. D., & Rupp, M. E. (2009). Coagulase-negative staphylococcal infections. *Infectious Disease Clinics of North America*, *23*(1), 73–98.

Sakinç, T., Michalski, N., Kleine, B., & Gatermann, S. G. (2009). The uropathogenic species Staphylococcus saprophyticus tolerates a high concentration of D-serine. *FEMS Microbiology Letters*, *299*(1), 60–64.

Sandle, T., Skinner, K., Sandle, J., Gebala, B., & Kothandaraman, P. (2013). Evaluation of the GEN III OmniLog®

ID System microbial identification system for the profiling of cleanroom bacteria. *European Journal of Parenteral & Pharmaceutical Sciences*, *18*(2), 44–50.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* , *30*(14), 2068–2069.

Suzuki, R., & Shimodaira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* , *22*(12), 1540–1542.

Tan, T. Y., Ng, S. Y., & Ng, W. X. (2006). Clinical significance of coagulase-negative staphylococci recovered from nonsterile sites. *Journal of Clinical Microbiology*, *44*(9), 3413–3414.

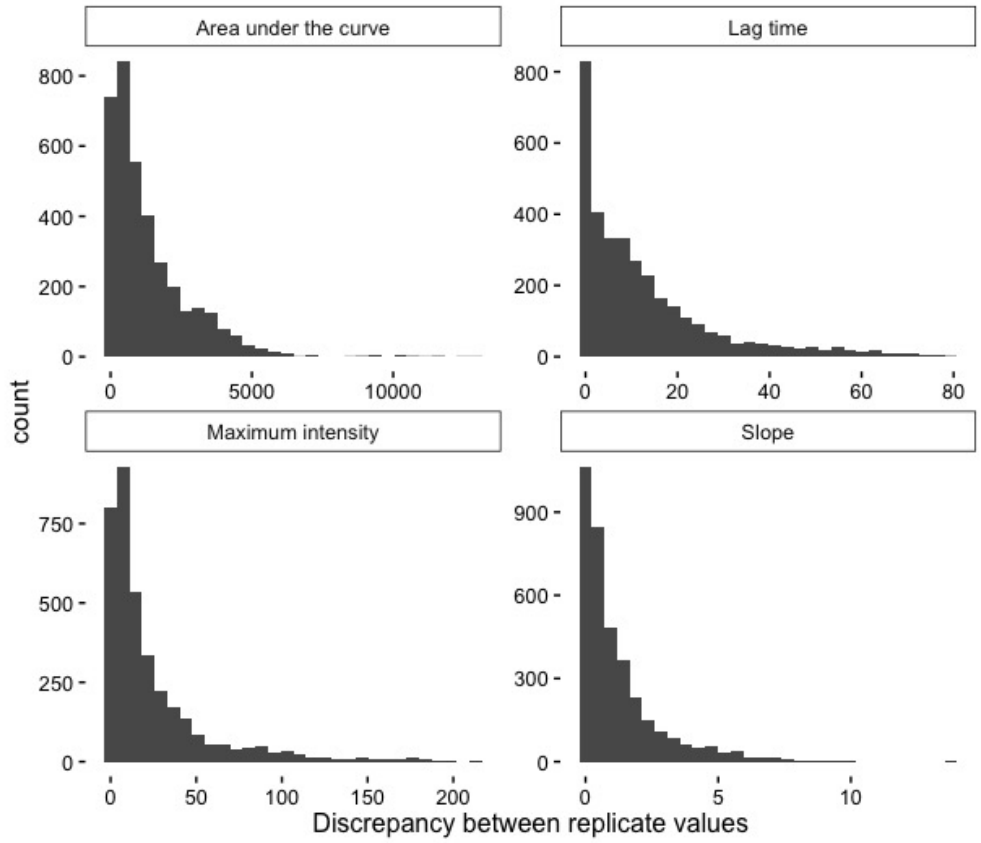
Vaas, L. A. I., Sikorski, J., Michael, V., Göker, M., & Klenk, H.-P. (2012). Visualization and curve-parameter estimation strategies for efficient exploration of phenotype microarray kinetics. *PloS One*, *7*(4), e34846.

Wragg, P., Randall, L., & Whatmore, A. M. (2014). Comparison of Biolog GEN III MicroStation semi-automated bacterial identification system with matrix-assisted laser desorption ionization-time of flight mass spectrometry and 16S ribosomal RNA gene sequencing for the identification of bacteria of veterinary interest. *Journal of Microbiological Methods*, *105*, 16–21.

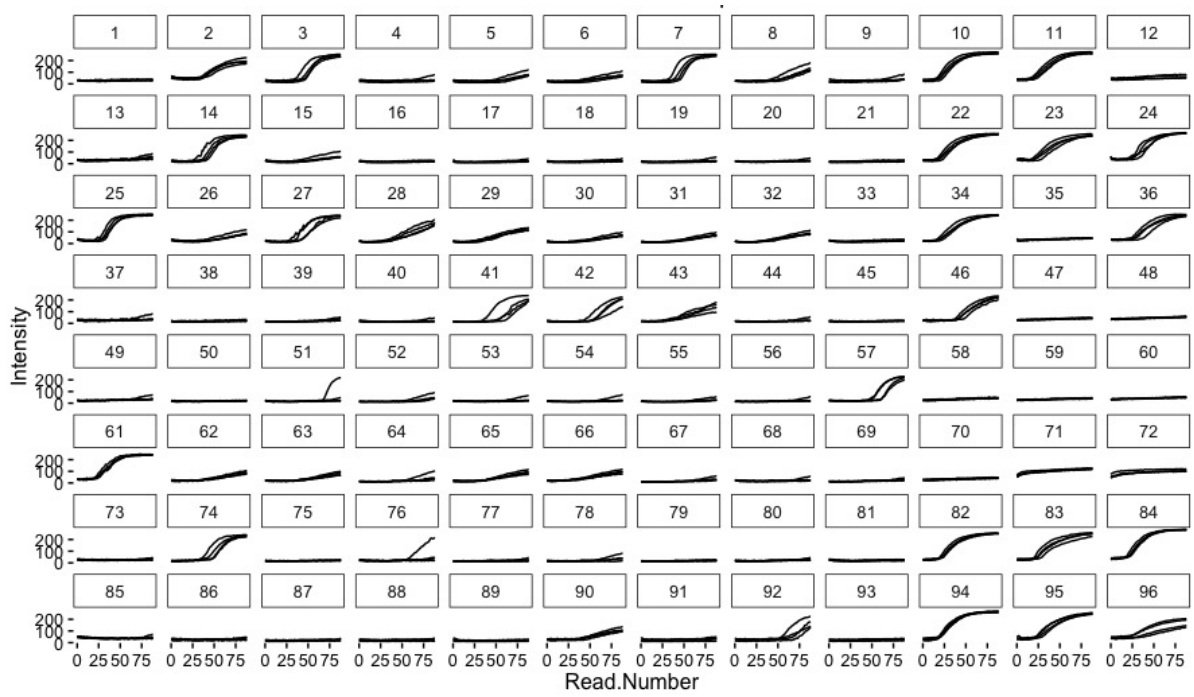
Supplementary Material

Supplementary Table 1 | Summary statistics for parameter estimates for replicates

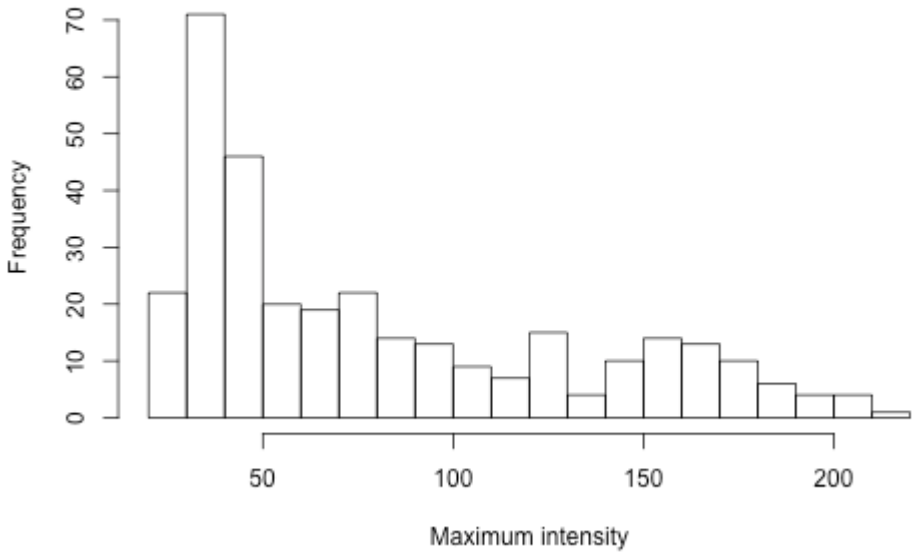
Parameter	Maximum value	Maximum within-strain range	Median within-strain range
Area under the curve	20803.57	12927.4	839.13
Lag time	86.71	79.40	6.84
Slope	18.98	13.77	0.65
Maximum intensity	311.67	213.19	11.86



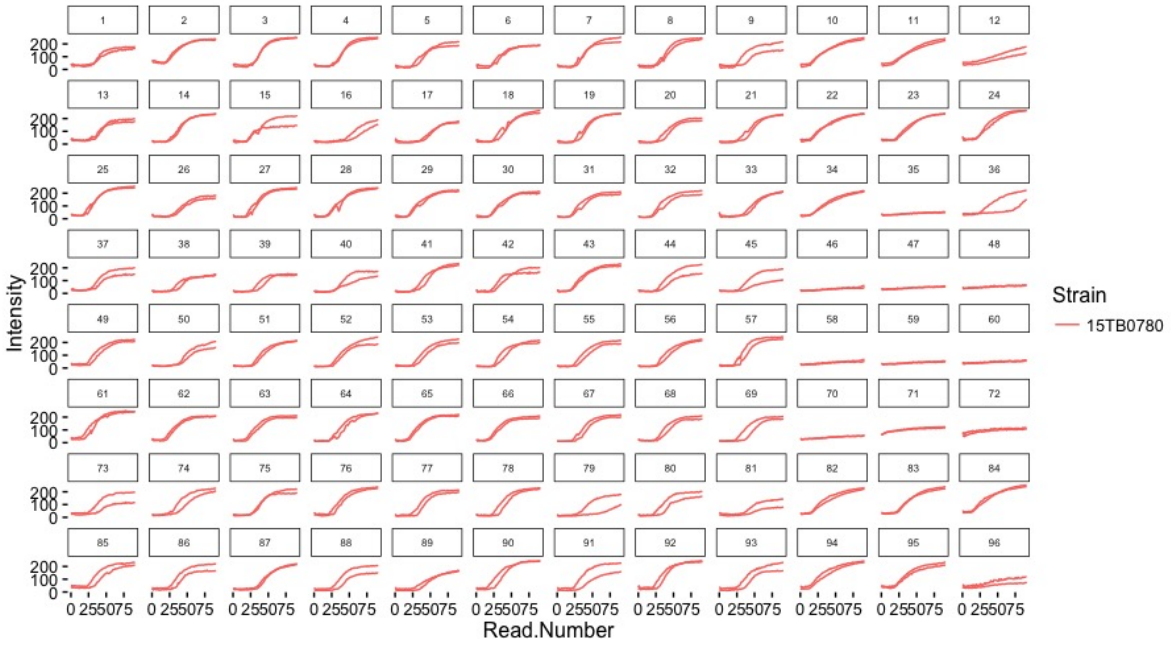
Supplementary Figure 1 | Plot of curve parameter estimate discrepancies between values



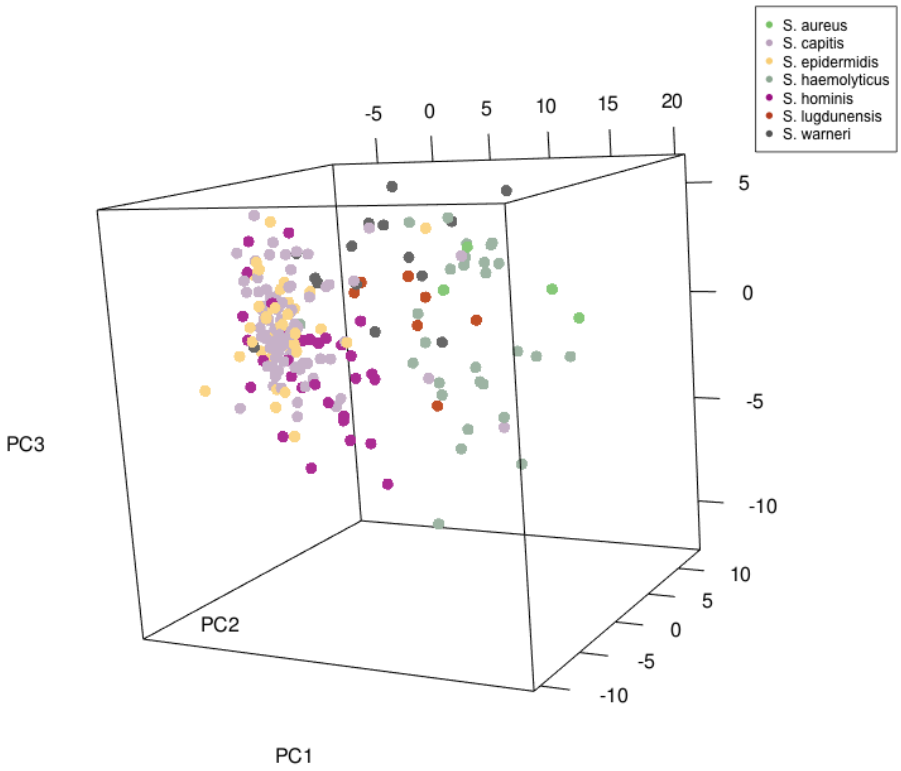
Supplementary Figure 2 | Consistency of curve shape for sample 27152, for which four replicates were run



Supplementary Figure 3 | Maximum intensity measurements in negative control well across all samples tested



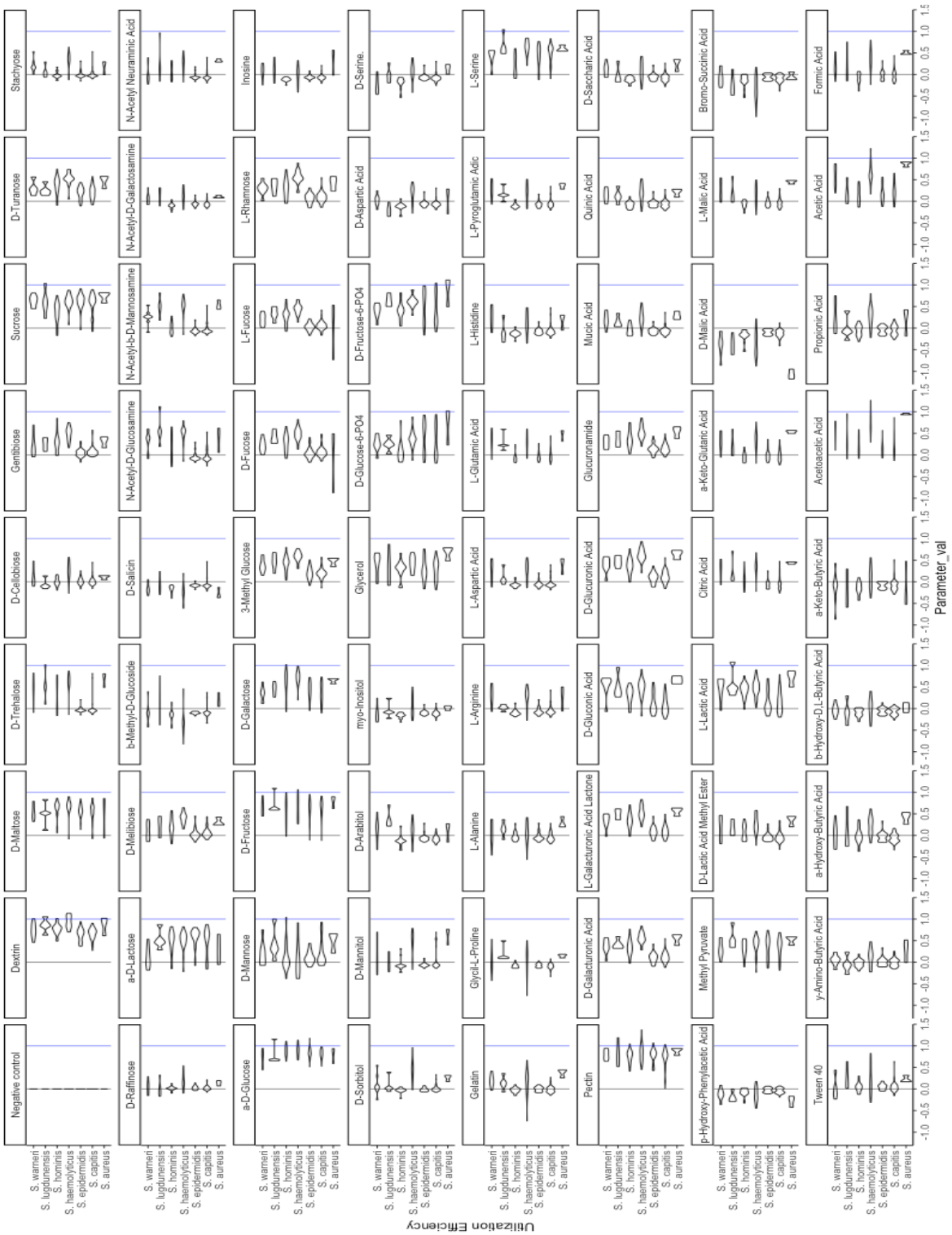
Supplementary Figure 4 | An example of a strain that showed reproducible growth in the negative control well and no growth in some of the inhibitory wells



Supplementary Figure 5 | PCA plotting of the AUC data identified structure in the data relating to species membership, however not enough to correctly distinguish species

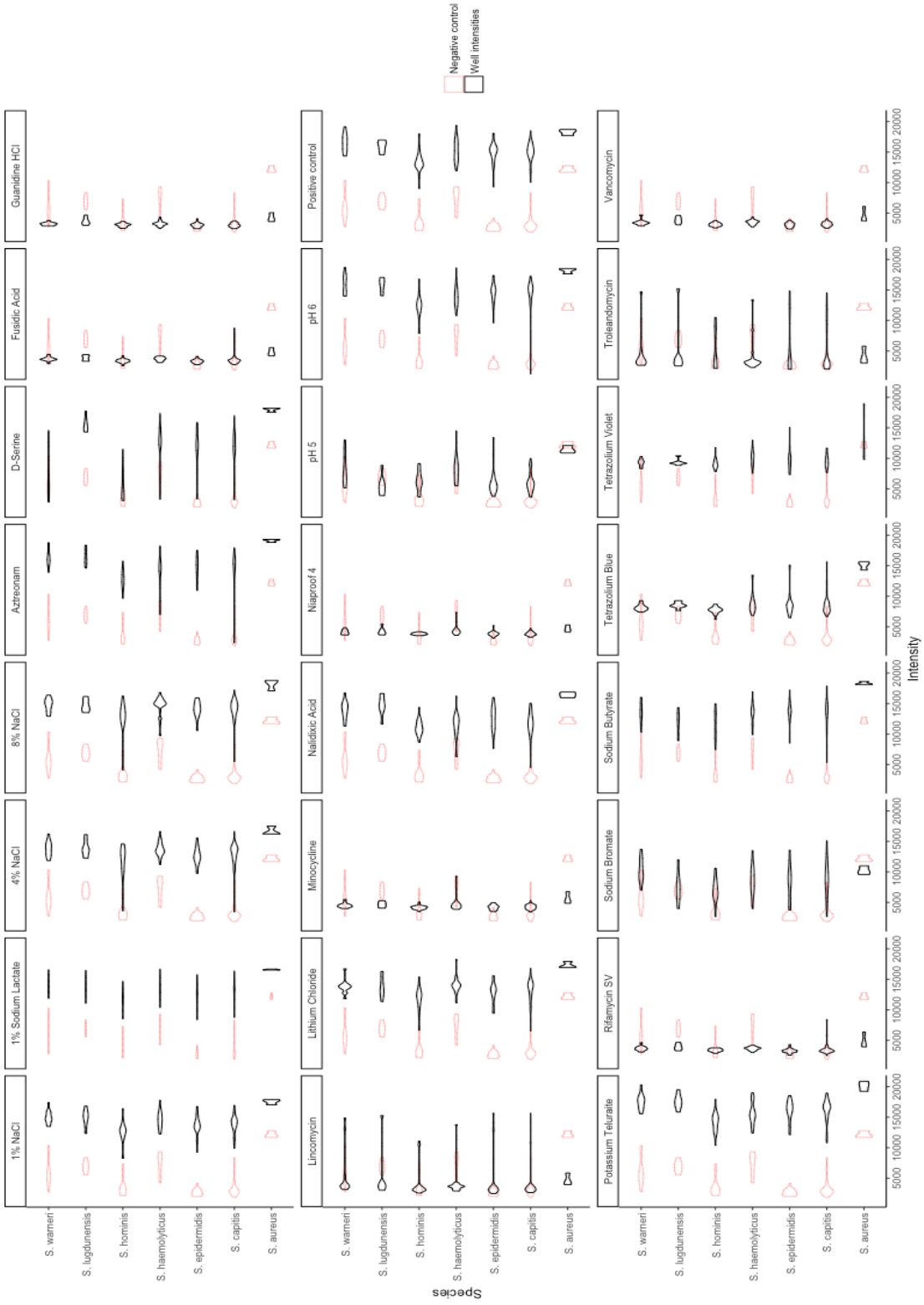
Supplementary Figure 6 | Areas under the curve for each sample in the study

Wells in the Biolog GEN III plate are clustered by similarity, and samples are grouped by species.



Supplementary Figure 7 | The distribution of area under the growth curve (AUC) values for samples from each species across Biolog GEN III carbohydrate sources

Growth equivalent to the negative control is shown with a black line and growth equivalent to the positive control is shown with a blue line.



Supplementary Figure 8 | The distribution of area under the growth curve (AUC) values for strains from each species across Biolog GEN III inhibitory substance wells

Growth in the negative control well is shown in pale red for comparison.

Chapter Eight | Discussion

Summary

This thesis has presented the development and expansion of the delta-bitscore (DBS) method of predicting functional degradation in protein coding sequences. The preceding chapters have shown the transition from its original conception as a method for performing pairwise comparisons of bacterial proteomes, to a highly scaleable approach designed with large comparative genomics studies in mind. It has also explored the processing and interpretation of large scale phenotypic characterizations of bacteria, a potential avenue for generating large amounts of phenotypic data to pair with DBS data for the investigation of genetic differences underlying metabolic differences. Figure 1 shows an outline of the different approaches that can be applied to analyzing DBS data.

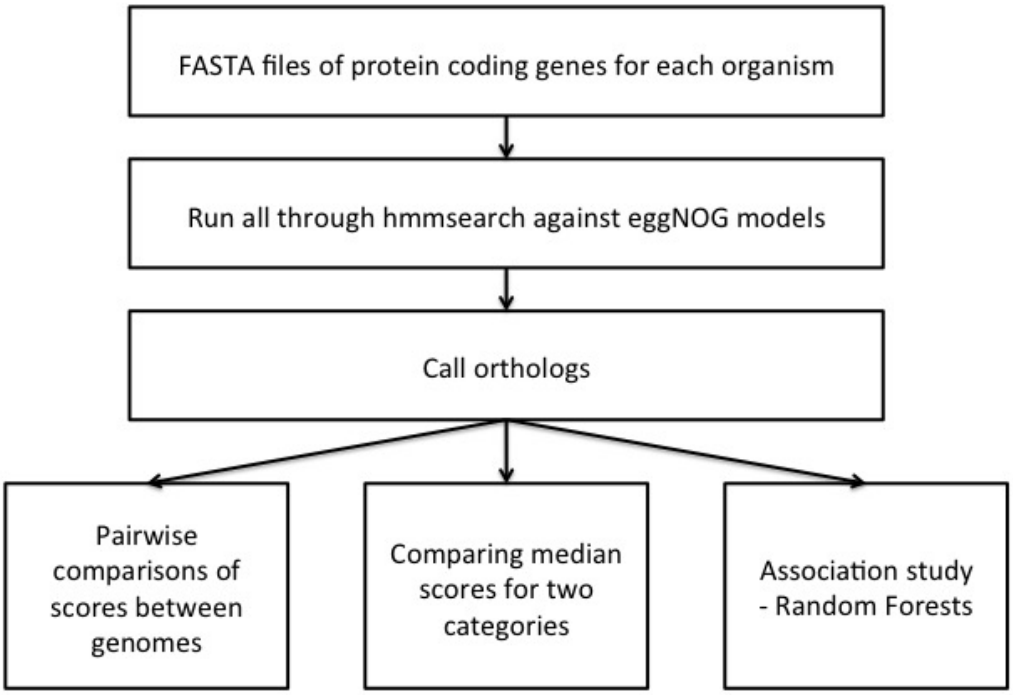


Figure 1 | An overview of the delta-bitscore workflow as presented in this thesis

In this thesis I have also demonstrated that loss of function mutations accumulate both as a result of genetic drift and also as reproducible signatures indicative of niche adaptation. These reproducible signatures can tell us about the differences in selective pressures and functional requirements of different environments. Genes could lose their function following a shift to a new niche for several reasons. They could be lost due to genetic streamlining, a particular form of Darwinian evolution (Holt et al., 2009; McClelland et al., 2004).

Alternatively, loss of genes could facilitate rapid metabolic rewiring to obtain greater metabolic efficiency in the new environment (Hottes et al., 2013). Mutations could inactivate antivirulence genes that impair fitness or threaten survival in the new environment (Maurelli, 2007). A final explanation could be that the loss of function is selectively neutral or even mildly deleterious, but the effect of purifying selection was strong enough in the old niche to purge these deleterious mutations, and not in the new niche, due to population bottlenecking (Zhou et al., 2014). While reasons for loss of function may vary, I expect that beneficial loss of function mutations will appear as stronger associations than those due to more neutral processes, giving us some ability to tell them apart.

In chapters Four to Six, I investigated the utility of DBS in detecting sequence differences associated with phenotype across a range of evolutionary distances. I expect the value of DBS will be greatest in identifying functionally important sequence differences in recently diverged clones, due to the fact that other methods designed to detect adaptive changes at the sequence level such as dN/dS do not perform well at short evolutionary distances (Rocha et al., 2006). However, the strength of the association signal we get from individual genes is weakest at short evolutionary distances, as observed in Chapter Four. Figure 2 below illustrates the difference in the strength of signal observed across different evolutionary distances, and demonstrates the greater overlap in scores seen in closely related strains of bacteria, the increasing separation of scores seen in serovars that have diverged a longer time ago, and finally the more discrete clustering of scores seen for inter-species comparisons.

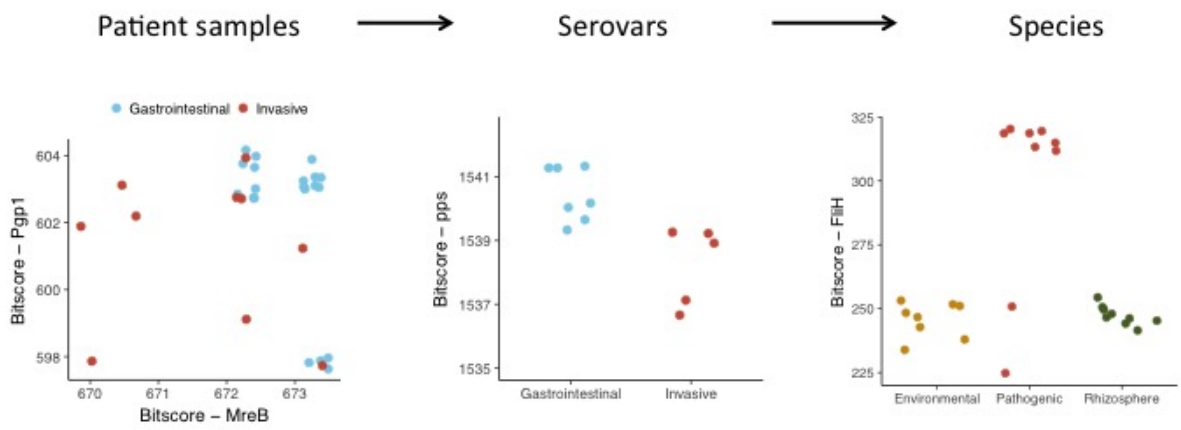


Figure 2 | Score distributions for key genes indicative of phenotype at the strain, serovar and species level

The “patient samples” plot is generated using key genes from the Chapter Four investigation of genetic markers of invasiveness in *Campylobacter*; the “serovar” plot is generated using a key gene identified in the Chapter Five study of genetic markers of invasiveness in *Salmonella*; the “species” plot is generated using a key gene identified in the Chapter Six study of genetic markers of pathogenicity in *Pseudomonas*.

This difference in both the degree of separation and the magnitude of the difference in bitscores across evolutionary distances placed some constraints on the development of the method. We initially wanted to assign a scoring threshold for identifying loss of function mutations, however this threshold would shift depending on evolutionary distance, and the relationship between the threshold and evolutionary distance may not have been straightforward. For the *E. coli* study performed in Chapter Three, we used a scoring threshold derived from benchmarking data in Chapter Two to identify loss of function mutations, due to the extremely short time scale of the evolution experiment and the controlled environmental conditions. However, in performing other comparative analyses, where strains are likely to live in different environments and encounter different pH and temperature ranges, we anticipated that the results derived from *in vitro* mutagenesis experiments performed under controlled conditions would not be directly applicable. The lack of a defined scoring threshold for analysis combined with the non-normal distribution of delta-bitscores we observed in our analyses restricted us to nonparametric statistical methods for identifying significant differences in bitscores. This created a challenge for analyzing bitscore differences in pairwise comparisons, and resulted in us selecting the most extreme values using an arbitrary cutoff percentage for the analysis in Chapter Two.

This was a major motivation for transitioning to comparisons of groups of organisms. Comparing distributions of scores presented a approach to identifying important differences in bitscore in a more sensible way, that was more likely to identify genuine functional differences in proteins, and differences in selective pressures encountered by bacteria leading different lifestyles. The group-based comparative method also increased the likelihood that the sequence changes we identified were genuinely associated with our phenotype of interest. Pairwise comparisons could only identify the most significant deviations in patterns of sequence variation in genes shared by two bacteria, but gave no evidence that these differences were relevant to our phenotype of interest. The newly developed group-based comparative tool gives us the opportunity to identify convergent changes in the same gene associated with a phenotype, which is strong evidence for adaptive change.

Application of DBS to microbial genome-wide association studies

Scaling of DBS to large comparative genomics datasets could result in an application essentially comparable to a microbial genome-wide association study (GWAS) (Falush, 2016; Lees & Bentley, 2016; Power et al., 2017). Microbial GWA studies are currently increasing in popularity and feasibility, and are of particular interest in cases where traits are difficult to study in the laboratory, such as host range or invasiveness, or where traits have a complex genetic basis (Falush, 2016). Generally speaking, a GWAS involves analysing a vast number of genetic variants, traditionally single nucleotide polymorphisms (SNPs), for association with a phenotype of interest. The concept of a GWAS has been around for over ten years in human genetics (Haines et al., 2005; Visscher et al., 2012; Welter et al., 2014), and the field has developed robust statistical and methodological approaches to ensuring results with minimal confounding due to biological factors (Korte & Farlow, 2013; McCarthy et al., 2008). In contrast, microbial GWASs are relatively new, with the first studies being reported in 2013 (Dutilh et al., 2013; Farhat et al., 2013; Sheppard et al., 2013). While we can take some methodological lessons from human GWAS, there are additional confounding factors that need to be addressed to perform a robust bacterial GWAS (Power et al., 2017).

Methodological considerations

The main issues in extending GWA studies to microbes stem from the differences in population structure and replication between humans and bacteria. An excellent review of the key differences can be found in (Power et al., 2017). I will discuss some of these issues, and how the method I have developed in this thesis can help to address them here.

Firstly, generating comparable variants across genomes presents a challenge in bacterial GWA studies that has so far been avoidable in human GWAS. In human GWAS, SNP chips are used to assay a common set of SNPs that typically achieve broad coverage of the genome (Spencer et al., 2009). Each SNP assayed on the chip generates a single, discrete value to test in an association study. In contrast, whole genome sequencing is used to generate data on genetic variation for microbial GWA studies. Genome rearrangements, recombination, and a high mutation rate will all make generating comparable data on variation difficult. Mapping reads to a reference genome is one solution to this problem, but does not account for gene loss/gain relative to the reference, nor does it deal well with genome rearrangements. Comparing “word” frequencies (short stretches of nucleotide sequence) is another solution that has been developed for bacterial GWAS that deals well with gene loss and gain events (Lees et al., 2016). However, single, functionally insignificant mutations in a genomic region corresponding to a given word will create an entirely new word, diluting signals of association between genomic regions and phenotype.

DBS addresses the issue of generating comparable data on genetic variation by comparing orthologous gene families. Comparing only orthologous groups, using protein family databases like Pfam or EggNOG has been suggested before, as it would improve the scalability of bacterial GWA studies (Dutilh et al., 2013). This approach is likely to result in the loss of many informative variants in intergenic regions, but allows the comparison of more genomes in a computationally efficient way, and allows the comparison of more diverse bacteria whose genomes are not easily aligned. The key advantage of incorporating DBS into a microbial GWAS approach is the ability to combine the effects of different variants in the same gene into a single, quantitative metric. If all SNPs within a gene have deleterious impacts to varying degrees, combining all of these into one variable reduces the number of test performed, increasing statistical power, and also allows us to include rare and even one-off variants.

Entire lineages can differ in phenotype, in which case an investigator needs to disentangle sequence differences that are related to lineage from sequence differences related to a particular trait (Power et al., 2017). As in human genetic studies where a trait may be more closely linked to one ethnic group, variants can be identified in the analysis that are more informative of ancestry than the biology of the trait of interest (Power et al., 2017). Because of this confounding effect, in large, highly structured collections of bacterial genomes we

must correct for population structure when performing association tests. In chapters Four and Five I have avoided the issue of population structure by using small, carefully selected datasets (although population structure presents a significant confounding effect in Chapter Six), but in scaling up the method this will be a critical issue that needs to be addressed. Correction for population structure can be approached by using mixed models to test for the association of a trait with a genetic variant, incorporating phylogenetic structure as a random effect in the model (Lippert et al., 2011). An extension of random forests that accounts for population structure has been introduced for human GWA studies (Stephan et al., 2015). This approach includes an additional random effect term in every split made by a tree, to search for the genetic marker that best separates classes after correcting for population structure. In moving to larger datasets, employing this random forest methodology with delta-bitscore data will be an effective approach to dealing with population structure.

Large scale genotype-phenotype associations

An important question in microbial GWAS using large numbers of samples is which phenotypes we are going to use for association studies and how confident we will be in the accuracy of their assignment to individual samples (Dutilh et al., 2013). In the case of invasive pathogens, identifying invasive potential can be complicated, especially for a large study. Just because an isolate was collected from someone not showing signs of invasive infection does not mean that the isolate does not have invasive potential in another host or under other circumstances. Similarly, if an isolate causes invasive infection in an immunocompromised individual, is it truly invasive or merely acting as an opportunistic invasive pathogen? One solution to this problem is to run phenotype arrays using easily testable, more objective phenotypes on the microbes that are being sequenced. If this is to be done, it is often as straightforward to do a single phenotypic test in high throughput as it is to do several, which could lead to extremely rich data sets for detecting associations.

A comparable approach has already been taken in the form of PhenoLink, which uses random forests paired with gene presence/absence data to find associations between gene content and phenotype (Bayjanov et al., 2012). PhenoLink uses sugar utilization assays and nitrogen dioxide production as measurable phenotypes. Biolog phenotype arrays would be a step up from this approach, giving us more phenotypic data on each isolate with a similar amount of time and effort. The multiple testing burden would then increase if we were to have large, multidimensional datasets with many phenotypes and many markers of genetic variation. For phenotypes such as those assayed by Biolog plates there are likely to be a subset of genes which show much stronger associations with ability to utilise substrates than

others, due to the somewhat binary nature of many substrate utilization phenotypes - if certain enzymes lose their function, substrate utilization could jump straight to zero. The large effect size of associated variants would increase our power to detect associations, potentially mitigating the effect of the large multiple testing burden.

Extensions of the method

Identifying conditional associations

Random forests could be particularly useful for identifying conditional associations between genes and phenotypes. Some examples include the contribution of a genetic variant to differences in phenotype only in a particular genetic background, gene-gene interactions and gene-environment interactions. We see a great illustration of the insight that can be gained by looking at epistasis in Tenaillon et al (2012), where the authors observed both positive and negative epistasis between beneficial mutations in long-term adaptation strategies to high temperatures in *E. coli*. The decision tree-based structure of random forests allows them to quantify the importance of variables under different contexts. If a gene is only informative given information on another gene, these two variables are more likely to co-occur in the same tree (Touw et al., 2013).

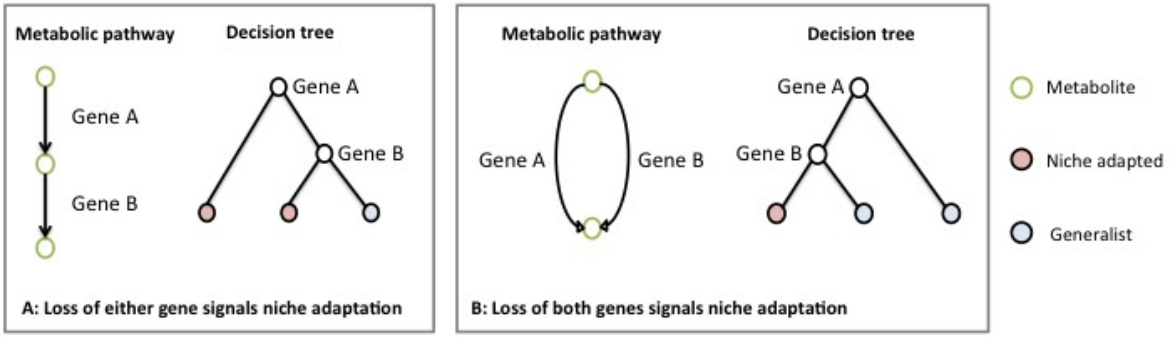


Figure 3 | Examples of conditional associations that could be captured by random forests

A: If two genes in a pathway contribute to metabolism of a nutrient not present in the new niche, the function of either gene could be stochastically lost during niche adaptation. This could be captured in a decision tree by interpreting loss of *either* gene A or gene B as a sign of niche adaptation. B: If two genes serve a redundant function metabolising a gene not present in the new niche, loss of both genes may signal niche adaptation. This could be captured in a decision tree by interpreting loss of gene A *and* gene B as a sign of niche adaptation.

An example that appears particularly relevant in the context of niche adaptation is gene-gene interactions relating to biochemical pathways important for niche adaptation. As Figure 3 illustrates, information not only on the co-occurrence of genes in a tree but also on their positions in the tree could give an indication of the relationship of the genes, e.g. whether they are both required for the same function or whether they perform redundant functions. There is no reason why environmental and patient variables could not also be incorporated into the model, for cases where immunocompromisation affects invasive ability, or where there is an interaction between lifestyles, for example host adaptation and invasiveness in *Salmonella enterica* (Rabsch et al., 2002).

Identifying misclassified samples

In training a classifier, some samples that have gone into the training data can be misclassified. For example, if a strain of bacteria was isolated from a patient exhibiting symptoms of a disease, but was not the causative factor, it may be mis-labelled as pathogenic. If these cases are in the minority, they could be identified by a random forest model using proximity scores for different samples. Proximity scores are calculated during model building for each sample, by running each sample through the random forest, and counting how many times two samples end up in the same terminal node of a tree (Touw et al., 2013). This metric gives an indication of how similar two samples are based on the key indicators of phenotype identified by the model. If a given sample is labelled as pathogenic, but appears more similar to nonpathogenic samples based on the key predictors in the model, this could give us cause to question the classification of the sample. Sometimes high similarity between pathogenic and nonpathogenic samples could indicate a recent switching of niche, meaning that not enough time has passed for sufficient loss of function mutations to accumulate for the sample to show the same genomic signatures seen in other niche-adapted isolates. Investigators would have to practise careful judgement in assessing class allocations based on experimental and clinical evidence compared to prediction by the model. Typically if evidence for a correct classification is strong, we would choose experimental evidence over the prediction by the model, but if our confidence in prior phenotype assignments is low, potentially mislabelled samples could be excluded from the model, tested further experimentally, or examined manually for genomic indicators of phenotype.

An iterative learning process

Collecting phenotypic data on large sets of sequenced bacteria is not always practical. There are large collections of genomes online already, however metadata associated with these strains can be inconsistent, and often do not include phenotypes of clinical significance. This means that often data collection for microbial GWAS has to be performed with such a study in mind, rather than using current publically available data. Another option is to iteratively train a model on a subset of well characterised data, then run the model on uncharacterised samples, classify them and add them to the training set. A similar approach has been applied to the building of profile hidden Markov models for remote homology detection and has been found to work well (Johnson et al., 2010). The potential for success of this approach remains to be tested, however it could be tested on large groups of phenotyped bacteria by taking a subset for model training, and testing the model on the remaining samples. The idea would be that there are a few sets of genes that lose their function earlier on, and could be detected with the original model. Other genes are lost more stochastically over time, so represent weaker signals. As the model captures more samples that display the highly replicable losses, power to identify weaker trends increases by increasing the sample size. There is potential for this iterative learning process to go awry, however our ability to experimentally test associations *in vivo* would allow us to verify that genes we have identified have a true impact on our phenotype of interest.

Could DBS work on nucleotide sequences?

A key limitation of DBS is that it only works on amino acid sequences of protein coding genes. We know that mutations in promoter sequences of bacteria can affect gene expression, allowing rapid adaptation to selective pressures (Corvec et al., 2002; Marvig et al., 2014). In addition, we know that many top SNPs identified in human GWAS studies have been intergenic (Li et al., 2016). Thus, it would be a vast improvement on DBS if it could effectively detect deleterious mutations in DNA sequence outside of coding regions. This is a further avenue for study, and would likely require a markedly different approach. Sequence conservation on the DNA level is less than that observed on the protein level, making alignment and interpretation of sequences difficult. In addition, DNA contains only four bases, compared to 20 canonical amino acids, making patterns of sequence variation less complex and harder to distinguish. This approach could be tested and tuned using saturation mutagenesis data (Baliga, 2001). If viable, the choice of which sequences to

include in each model, and the appropriate evolutionary distance or sequence identity cutoff to use would need to be explored.

Open questions

Interpreting low variance in bitscores

In the studies performed in this thesis, I have been assuming that low variance in bitscore for a gene for a particular group indicates high conservation of function due to strong selection on that gene in a particular niche, while this may not actually be the case. David et al (2016) found in an investigation of *Legionella* associated with clinical infections a range of nucleotides showing lower diversity in pathogens than in non pathogens. This was eventually traced back to recombination events, where it appeared that the same region had been acquired by several phylogenetically distinct pathogenic lineages of *Legionella*. This may also indicate that a region is important for pathogenicity, in a scenario where acquisition of this region of DNA confers pathogenic potential or advantage. Further testing for recombination events using other software would resolve this question. A more straightforward solution may be available in some cases, where while the distribution of delta bitscores is much tighter in one group than in the other, the number of nonsynonymous changes is actually quite similar. This indicates that a similar amount of sequence change has occurred over time in bacteria from both niches, but deleterious mutations have been avoided in one niche.

Are the effects of mutations in a protein additive?

A key assumption behind DBS is that it is appropriate to sum the individual scores for residues in a protein that match the profile HMM, to get an overall indication of how likely individual deviations of residues from the predicted sequence constraints on a protein are to cause a loss of protein function. The benchmarking datasets used in Chapter Two use data on single amino acid substitutions, so give us no indication of how DBS would perform on combinations of mutations occurring in the same protein. Are the effects of individual mutations on protein function truly additive? Fortunately, experimental evidence suggests that most mutations do have additive effects, with a few exceptions mostly involving physical contacts between residues (Reetz, 2013; Skinner & Terwilliger, 1996). Modelling non-additive effects would be an additional challenge for the method, and would require an extension of profile hidden Markov models, possibly in a way similar to covariance models in ncRNAs (Eddy & Durbin, 1994) or a covarion model (Wang et al., 2009).

Concluding statement

The delta-bitscore metric as outlined in this thesis presents a novel method of identifying functionally significant sequence variation in bacterial genomes, and the statistical and machine learning approaches employed present opportunities to use this metric to identify sequence variation associated with phenotype. The weighting of sequence variants according to their likely functional impact, and the combination of information on sequence variation within the same gene to a single, comparable metric represent key improvements in increasing the power of statistical methods to detect true associations. These advances come at a critical time, when our approach to genome-wide association studies in bacteria is still being developed. A challenge in the interpretation of GWAS results is the large number of associations that result from analyses and our ability to draw meaningful conclusions from these results. The value of human GWA studies has been questioned, with major criticisms including the lack of biological insight that has been generated by them and the spurious nature of the results (Visscher et al., 2012). Tying variants to specific proteins and increasing confidence that associations identified are related to measurable differences in protein function allows for greater interpretability of results and a greater probability that experimental validation of associations will result in measurable differences in phenotype. Arguments can be made about whether it is preferable to identify a small number of confident, testable associations or whether it is better to identify the most comprehensive set of associations possible while still maintaining statistical rigour. I believe the method proposed in this thesis draws a fair balance between the two approaches, and offers a novel approach to bacterial GWA studies that has unique strengths when compared to other currently available methods.

References

Baliga, N. S. (2001). Promoter analysis by saturation mutagenesis. *Biological Procedures Online*, *3*, 64–69.

Bayjanov, J. R., Molenaar, D., Tzeneva, V., Siezen, R. J., & van Hijum, S. A. F. T. (2012). PhenoLink--a web-tool for linking phenotype to ~omics data for bacteria: application to gene-trait matching for Lactobacillus plantarum strains. *BMC Genomics*, *13*, 170.

Corvec, S., Caroff, N., Espaze, E., Marraillac, J., & Reynaud, A. (2002). -11 Mutation in the ampC Promoter Increasing Resistance to -Lactams in a Clinical Escherichia coli Strain. *Antimicrobial Agents and Chemotherapy*, *46*(10), 3265–3267.

David, S., Rusniok, C., Mentasti, M., Gomez-Valero, L., Harris, S. R., Lechat, P., … Buchrieser, C. (2016). Multiple major disease-associated clones of Legionella pneumophila have emerged recently and independently. *Genome Research*, *26*(11), 1555–1564.

Dutilh, B. E., Backus, L., Edwards, R. A., Wels, M., Bayjanov, J. R., & van Hijum, S. A. F. T. (2013). Explaining microbial phenotypes on a genomic scale: GWAS for microbes. *Briefings in Functional Genomics*, *12*(4), 366–380.

Eddy, S. R., & Durbin, R. (1994). RNA sequence analysis using covariance models. *Nucleic Acids Research*, *22*(11), 2079–2088.

Falush, D. (2016). Bacterial genomics: Microbial GWAS coming of age. *Nature Microbiology*, *1*, 16059.

Farhat, M. R., Shapiro, B. J., Kieser, K. J., Sultana, R., Jacobson, K. R., Victor, T. C., … Murray, M. (2013). Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium

tuberculosis. *Nature Genetics*, *45*(10), 1183–1189.
Haines, J. L., Hauser, M. A., Schmidt, S., Scott, W. K., Olson, L. M., Gallins, P., … Pericak-Vance, M. A. (2005). Complement factor H variant increases the risk of age-related macular degeneration. *Science*, *308*(5720), 419–421.
Holt, K. E., Thomson, N. R., Wain, J., Langridge, G. C., Hasan, R., Bhutta, Z. A., … Parkhill, J. (2009). Pseudogene accumulation in the evolutionary histories of Salmonella enterica serovars Paratyphi A and Typhi. *BMC Genomics*, *10*, 36.
Hottes, A. K., Freddolino, P. L., Khare, A., Donnell, Z. N., Liu, J. C., & Tavazoie, S. (2013). Bacterial adaptation through loss of function. *PLoS Genetics*, *9*(7), e1003617.
Johnson, L. S., Eddy, S. R., & Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, *11*, 431.
Korte, A., & Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods*, *9*, 29.
Lees, J. A., & Bentley, S. D. (2016). Bacterial GWAS: not just gilding the lily. *Nature Reviews Microbiology*, *14*(7), 406.
Lees, J. A., Vehkala, M., Välimäki, N., Harris, S. R., Chewapreecha, C., Croucher, N. J., … Corander, J. (2016). Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nature Communications*, *7*, 12797.
Li, H., Achour, I., Bastarache, L., Berghout, J., Gardeux, V., Li, J., … Lussier, Y. A. (2016). Integrative genomics analyses unveil downstream biological effectors of disease-specific polymorphisms buried in intergenic regions. *NPJ Genomic Medicine*, *1*. https://doi.org/10.1038/npjgenmed.2016.6
Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., & Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nature Methods*, *8*(10), 833–835.
Marvig, R. L., Damkiaer, S., Khademi, S. M. H., Markussen, T. M., Molin, S., & Jelsbak, L. (2014). Within-Host Evolution of Pseudomonas aeruginosa Reveals Adaptation toward Iron Acquisition from Hemoglobin. *mBio*, *5*(3), e00966–14–e00966–14.
Maurelli, A. T. (2007). Black holes, antivirulence genes, and gene inactivation in the evolution of bacterial pathogens. *FEMS Microbiology Letters*, *267*(1), 1–8.
McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., & Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, *9*(5), 356–369.
McClelland, M., Sanderson, K. E., Clifton, S. W., Latreille, P., Porwollik, S., Sabo, A., … Wilson, R. K. (2004). Comparison of genome degradation in Paratyphi A and Typhi, human-restricted serovars of Salmonella enterica that cause typhoid. *Nature Genetics*, *36*(12), 1268–1274.
Power, R. A., Parkhill, J., & de Oliveira, T. (2017). Microbial genome-wide association studies: lessons from human GWAS. *Nature Reviews Genetics*, *18*(1), 41–50.
Rabsch, W., Andrews, H. L., Kingsley, R. A., Prager, R., Tschäpe, H., Adams, L. G., & Bäumler, A. J. (2002). Salmonella enterica serotype Typhimurium and its host-adapted variants. *Infection and Immunity*, *70*(5), 2249–2255.
Reetz, M. T. (2013). The importance of additive and non-additive mutational effects in protein engineering. *Angewandte Chemie*, *52*(10), 2658–2666.
Rocha, E. P. C., Smith, J. M., Hurst, L. D., Holden, M. T. G., Cooper, J. E., Smith, N. H., & Feil, E. J. (2006). Comparisons of dN/dS are time dependent for closely related bacterial genomes. *Journal of Theoretical Biology*, *239*(2), 226–235.
Sheppard, S. K., Didelot, X., Meric, G., Torralbo, A., Jolley, K. A., Kelly, D. J., … Falush, D. (2013). Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(29), 11923–11927.
Skinner, M. M., & Terwilliger, T. C. (1996). Potential use of additivity of mutational effects in simplifying protein engineering. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(20), 10753–10757.
Spencer, C. C. A., Su, Z., Donnelly, P., & Marchini, J. (2009). Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genetics*, *5*(5), e1000477.
Stephan, J., Stegle, O., & Beyer, A. (2015). A random forest approach to capture genetic effects in the presence of population structure. *Nature Communications*, *6*, 7432.
Tenaillon, O., Rodríguez-Verdugo, A., Gaut, R. L., McDonald, P., Bennett, A. F., Long, A. D., & Gaut, B. S. (2012). The molecular diversity of adaptive convergence. *Science*, *335*(6067), 457–461.
Touw, W. G., Bayjanov, J. R., Overmars, L., Backus, L., Boekhorst, J., Wels, M., & van Hijum, S. A. F. T. (2013). Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? *Briefings in Bioinformatics*, *14*(3), 315–326.
Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *American Journal of Human Genetics*, *90*(1), 7–24.
Wang, H.-C., Susko, E., & Roger, A. J. (2009). PROCOV: maximum likelihood estimation of protein phylogeny under covarion models and site-specific covarion pattern analysis. *BMC Evolutionary Biology*, *9*, 225.
Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., … Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, *42*(Database issue), D1001–6.

Zhou, Z., McCann, A., Weill, F.-X., Blin, C., Nair, S., Wain, J., … Achtman, M. (2014). Transient Darwinian selection in Salmonella enterica serovar Paratyphi A during 450 years of global spread of enteric fever. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(33), 12199–12204.